DNA in chromatin :

how to extract structural, dynamical and functional information from the analysis of genomic sequences using space-scale wavelet techniques

Alain Arneodo

Laboratoire de Physique, Ecole Normale Supérieure de Lyon
46 allée d'Italie, 69364 Lyon Cedex 07, FRANCE

Françoise Argoul
Benjamin Audit
Samuel Nicolay — ENS de Lyon, France
Edward-Benedict Brodie of Brodie

Cédric Vaillant — EPF Lausanne, Switzerland

Marie Touchon
Yves d'Aubenton-Carafa — CGM, Gif-sur-Yvette, France
Claude Thermes

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE
07 JAN 2005

2. REPORT TYPE
N/A

3. DATES COVERED
-

4. TITLE AND SUBTITLE
DNA in chromatin:how to extract structural, dynamical and functional information from the analysis of genomic sequences using space-scale wavelet techniques

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Laboratoire de Physique,EcoleNormale Supérieure de Lyon46 allée dItalie, 69364 Lyon Cedex 07, FRANCE

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001750, Wavelets and Multifractal Analysis (WAMA) Workshop held on 19-31 July 2004., The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:

a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
unclassified	**unclassified**	**unclassified**	**UU**	**60**	

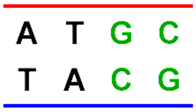
DESOXYRIBONUCLEIC ACID

A FEW HISTORICAL LANDMARKS

1869 Miescher isolates DNA

1944 DNA carries the genetic information (Avery)

1953 The double helix structure of DNA is discovered by Watson and Crick

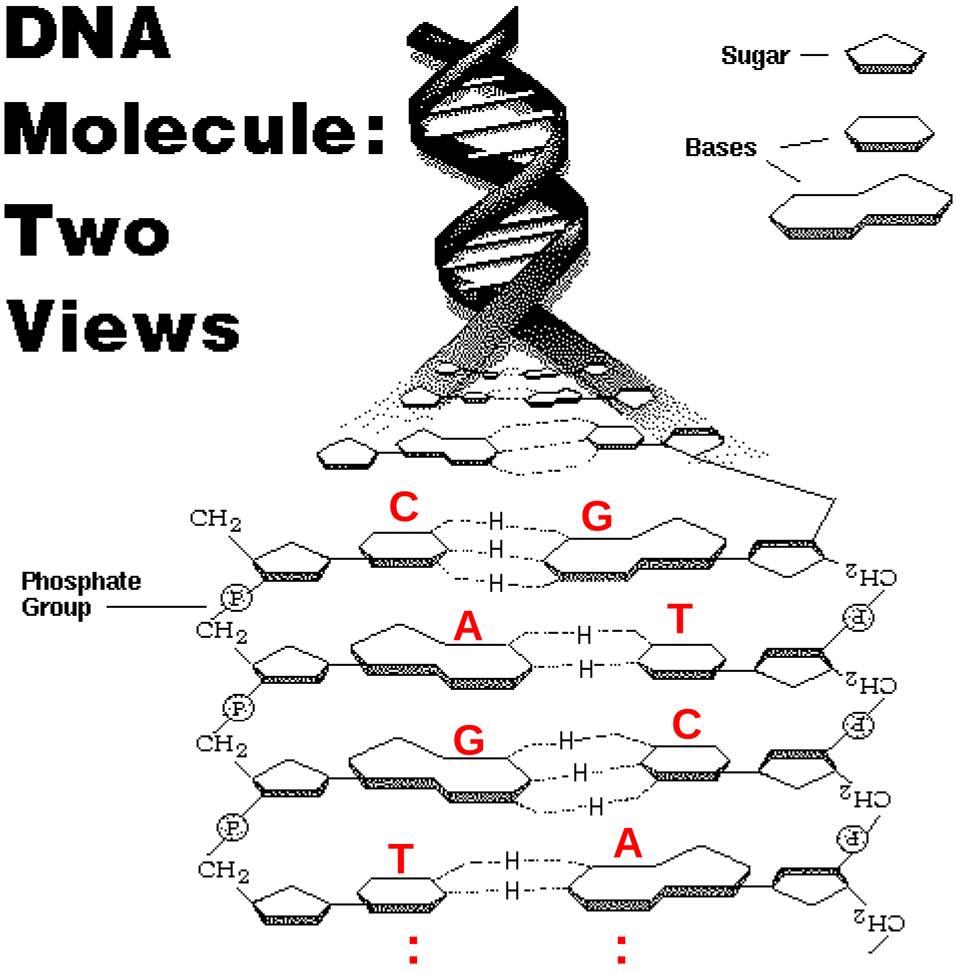


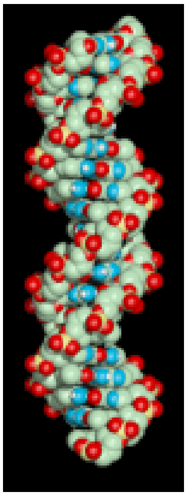
→ a simple model for the transmission of the genetic information

1966 Niremberg, Ochoa and Khorana elucidate the genetic code
→ DNA codes for proteins

codon	ATG	GCG	ACG	. . .	GCC	GTG	TAA
amino acid	Met	Ala	Thr	· · ·	Ala	Val	
	start						stop

DeoxyriboNucleic Acid





- Double helix macromolecule

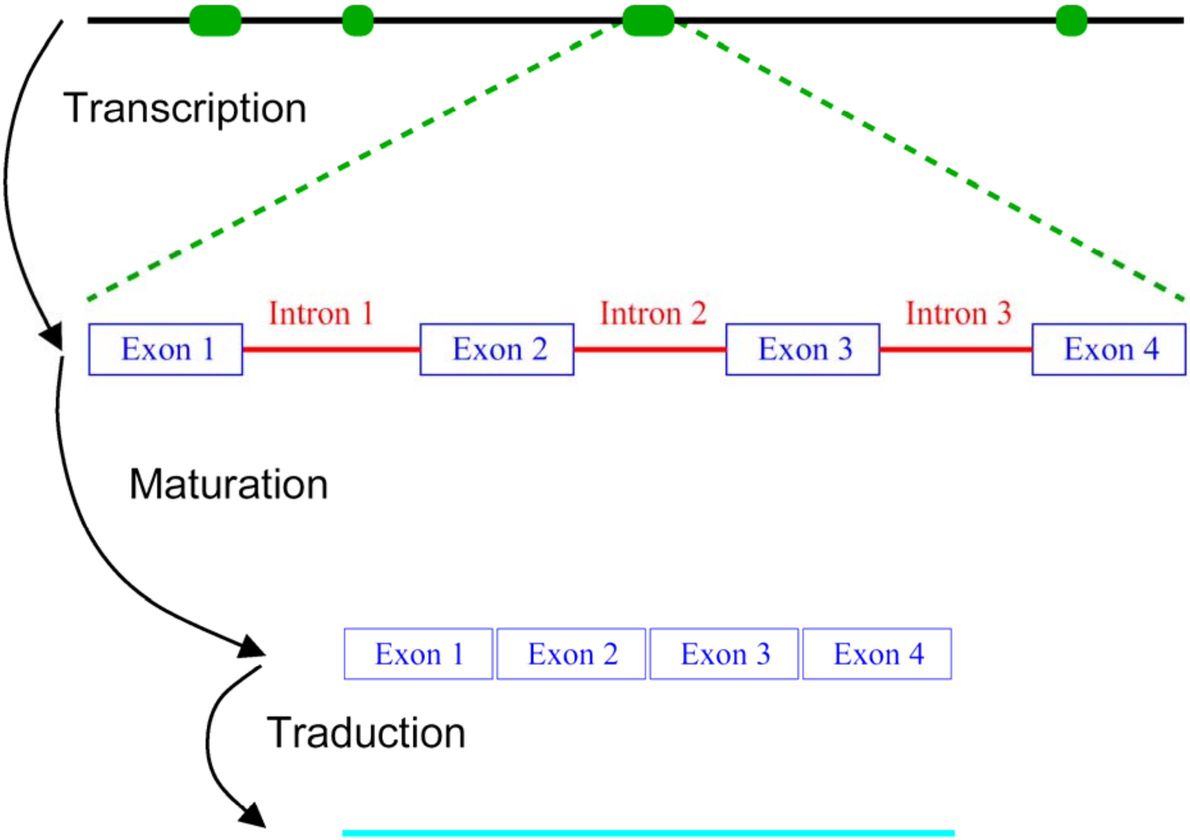
- Each strand consists of an oriented sequence of four possible nucleotides:

  Adenine, Thymine, Guanine & Cytosine

- Complementary strands:

  [A]=[T] & [G]=[C] over the sum of both strands

Organization of the human genome



Transcription

Exon 1 Intron 1 Exon 2 Intron 2 Exon 3 Intron 3 Exon 4

Maturation

Exon 1 Exon 2 Exon 3 Exon 4

Traduction

23 Chromosomes
$L \sim$ 100Mbp.

Genes ($\sim$ 20%)
$L \sim$ 10kbp.

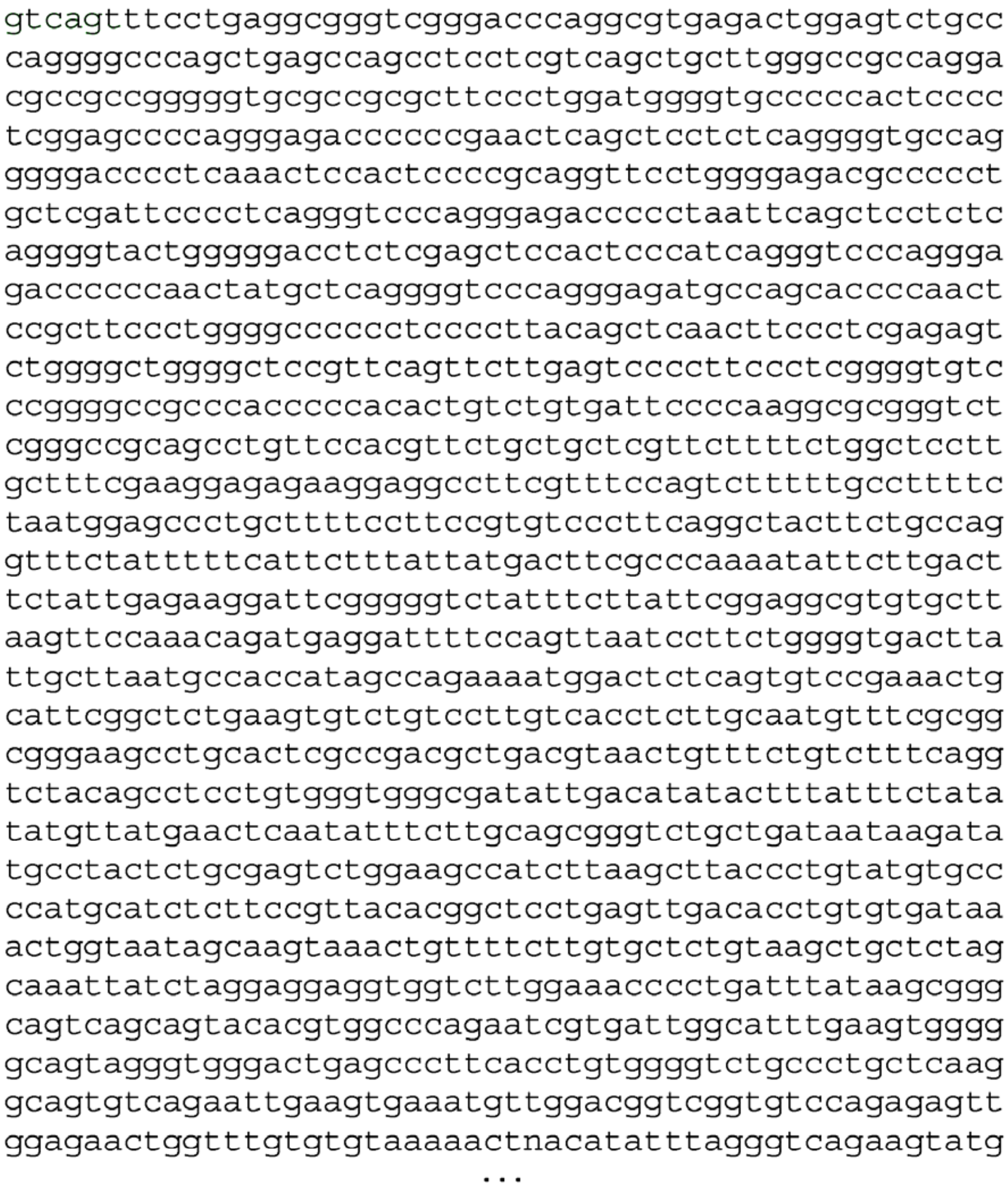
Introns
(INTervening seq.)
$L \sim$ 1kbp.

Exons
(EXpressed seq.) $\Rightarrow$
$L \sim$ 150bp.

Proteins
$L \sim$ 500AA.

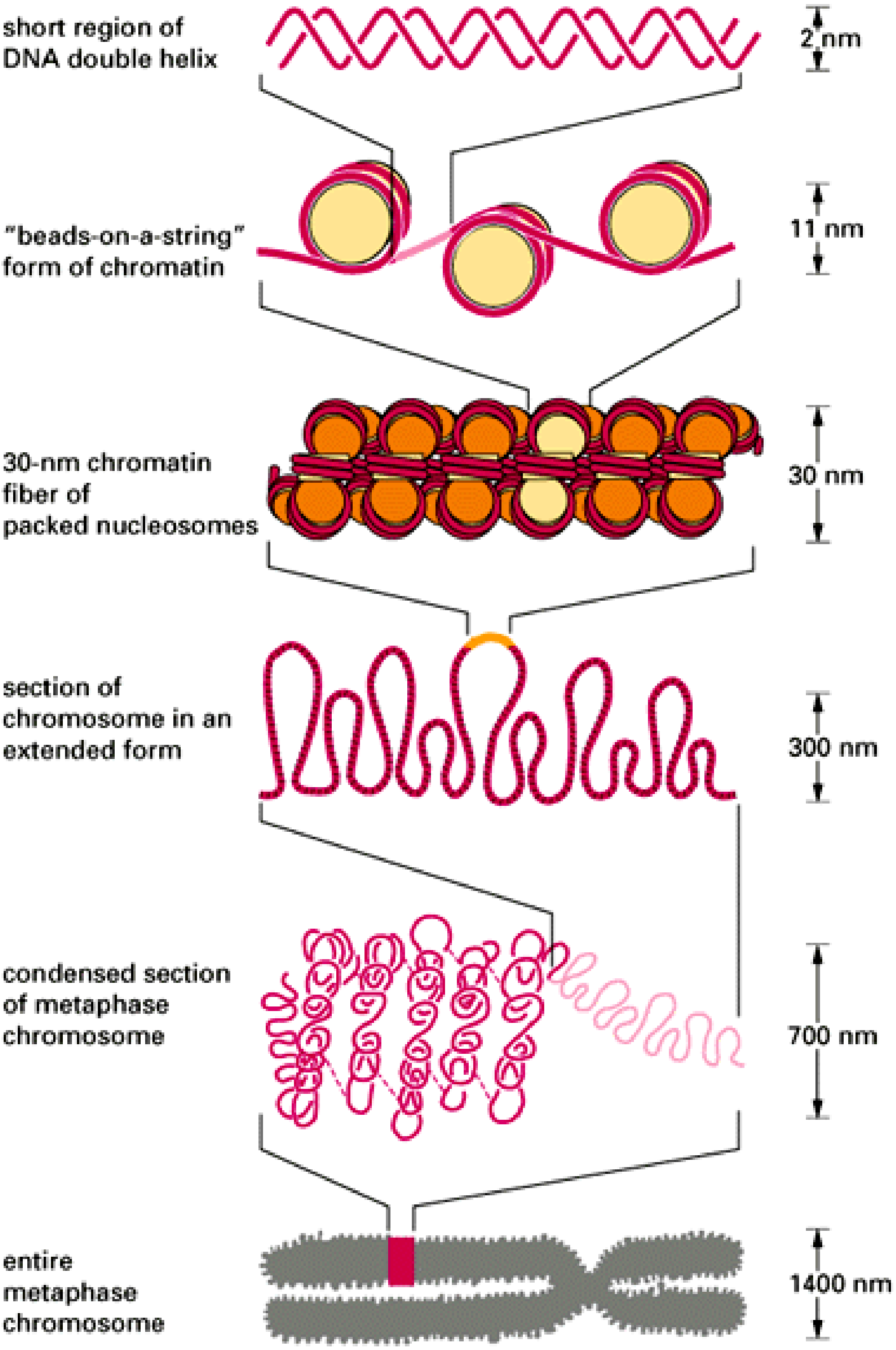
Non genic DNA

Sequencing projects result in 4 letter texts :



```
gtcagtttcctgaggcgggtcgggacccaggcgtgagactggagtctgcc
caggggcccagctgagccagcctcctcgtcagctgcttgggccgccagga
cgccgccgggggtgcgccgcgcttccctggatggggtgcccccactcccc
tcggagccccagggagacccccccgaactcagctcctctcaggggtgccag
ggggacccctcaaactccactccccgcaggttcctggggagacgcccccct
gctcgattcccctcagggtcccagggagacccccctaattcagctcctctc
aggggtactgggggacctctcgagctccactcccatcagggtcccaggga
gacccccccaactatgctcaggggtcccagggagatgccagcaccccaact
ccgcttccctgggggccccccctccccttacagctcaacttccctcgagagt
ctggggctggggctccgttcagttcttgagtccccttccctcggggtgtc
ccgggggccgcccaccccccacactgtctgtgattccccaaggcgcgggtct
cgggccgcagcctgttccacgttctgctgctcgttcttttctggctcctt
gctttcgaaggagagaaggaggccttcgtttccagtcttttttgcctttttc
taatggagccctgcttttccttccgtgtcccttcaggctacttctgccag
gtttctatttttcattctttattatgacttcgcccaaaatattcttgact
tctattgagaaggattcgggggtctatttcttattcggaggcgtgtgctt
aagttccaaacagatgaggattttccagttaatccttctggggtgactta
ttgcttaatgccaccatagccagaaaatggactctcagtgtccgaaactg
cattcggctctgaagtgtctgtccttgtcacctcttgcaatgtttcgcgg
cgggaagcctgcactcgccgacgctgacgtaactgtttctgtctttcagg
tctacagcctcctgtgggtgggcgatattgacatatactttatttctata
tatgttatgaactcaatatttcttgcagcgggtctgctgataataagata
tgcctactctgcgagtctggaagccatcttaagcttaccctgtatgtgcc
ccatgcatctcttccgttacacggctcctgagttgacacctgtgtgataa
actggtaatagcaagtaaactgttttcttgtgctctgtaagctgctctag
caaattatctaggaggaggtggtcttggaaacccctgatttataagcggg
cagtcagcagtacacgtggcccagaatcgtgattggcatttgaagtgggg
gcagtagggtgggactgagcccttcacctgtggggtctgccctgctcaag
gcagtgtcagaattgaagtgaaatgttggacggtcggtgtccagagagtt
ggagaactggtttgtgtgtaaaaactnacatatttagggtcagaagtatg
                         ...
```

HIERARCHICAL STRUCTURE OF EUCARYOTIC DNA


short region of DNA double helix
2 nm
"beads-on-a-string" form of chromatin
11 nm
30-nm chromatin fiber of packed nucleosomes
30 nm
section of chromosome in an extended form
300 nm
condensed section of metaphase chromosome
700 nm
entire metaphase chromosome
1400 nm

NET RESULT : EACH DNA MOLECULE HAS BEEN PACKAGED INTO A MITOTIC CHROMOSOME THAT IS 50.000x SHORTER THAN ITS EXTENDED LENGTH

DIFFERENT WAYS TO READ THE TEXT

I. "Classical" reading

- Looking for patterns
  - Genes, introns, exons detection
  - Splicing sites, promoters, replication origins recognition
- Characterizing repetitions
  - Tandem, interspersed repeats
  - Oligonucleotide usage
- Using methods such as
  - Hidden Markov chains
  - Fourier transform
  - Dot-plot matrices and recurrence plots

INVARIANCE UNDER TRANSLATION

II. The physicist reading

- Hypothesis: The DNA text results from a stochastic process :

  ACGTTCGAT ?

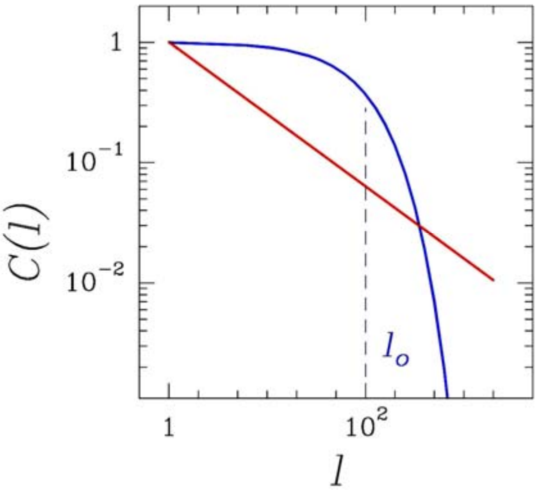
- Question: The choice of the next nucleotide :

  i. Depends on a finite number (l_o) of the previous trials
  → Short range correlations and exponential decay of the correlation function:

  $$C(l) \propto \exp(-l/l_o)$$

  ii. Depends on all the previous nucleotides
  → Long range correlations and power law decay of the correlation function:

  $$C(l) \propto l^{-\kappa}$$



INVARIANCE UNDER DILATATION

DNA WALK REPRESENTATION (PENG *et al.* 92)

1. Each nucleotide is associated to a numerical value (A to a, T to t, G to g and C to c).

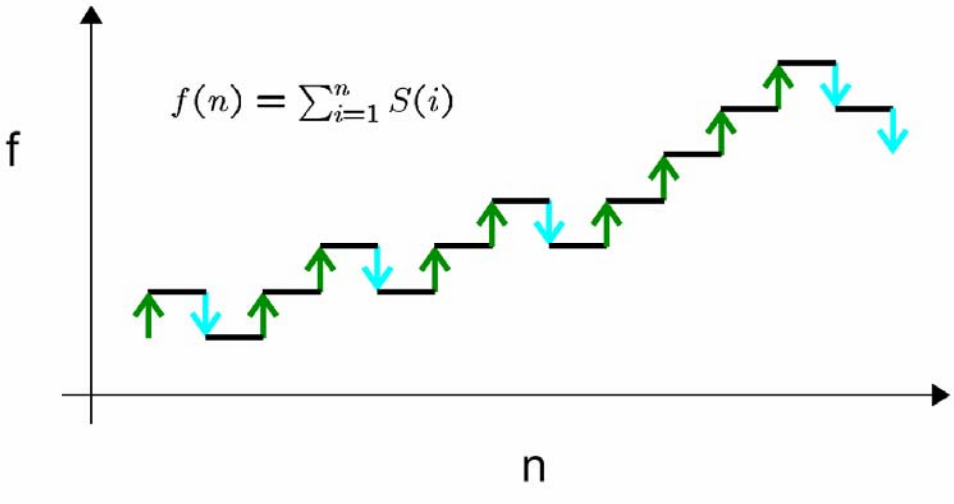
   purine-pyrimidine : $a = g = 1$ and $t = c = -1$
   weak-strong : $a = t = 1$ and $g = c = -1$
   amino-keto : $a = c = 1$ and $t = g = -1$

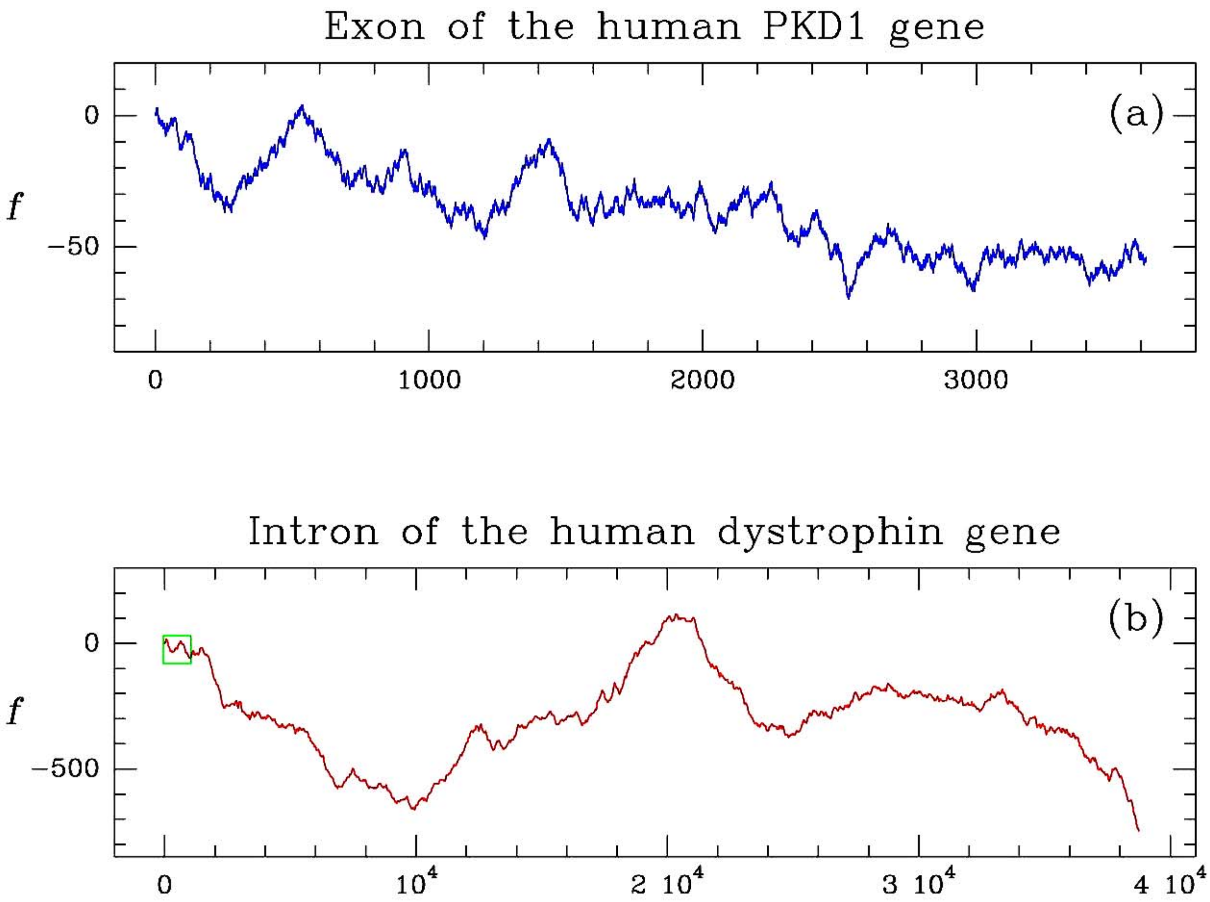
   A-non A : $a = 1$ and $t = g = c = -1/3$
   T-non T : $t = 1$ and $a = g = c = -1/3$
   G-non G : $g = 1$ and $a = t = c = -1/3$
   C-non C : $c = 1$ and $a = t = g = -1/3$

2. Suppose you have a walker on the line. The value associated to the i^{th} nucleotide defines the i^{th} step $S(i)$ of the walker

Example using the purine (↑) pyrimidine (↓) distinction :

A T G G C G A C G A A G C T
↑ ↓ ↑ ↑ ↓ ↑ ↑ ↓ ↑ ↑ ↑ ↑ ↓ ↓

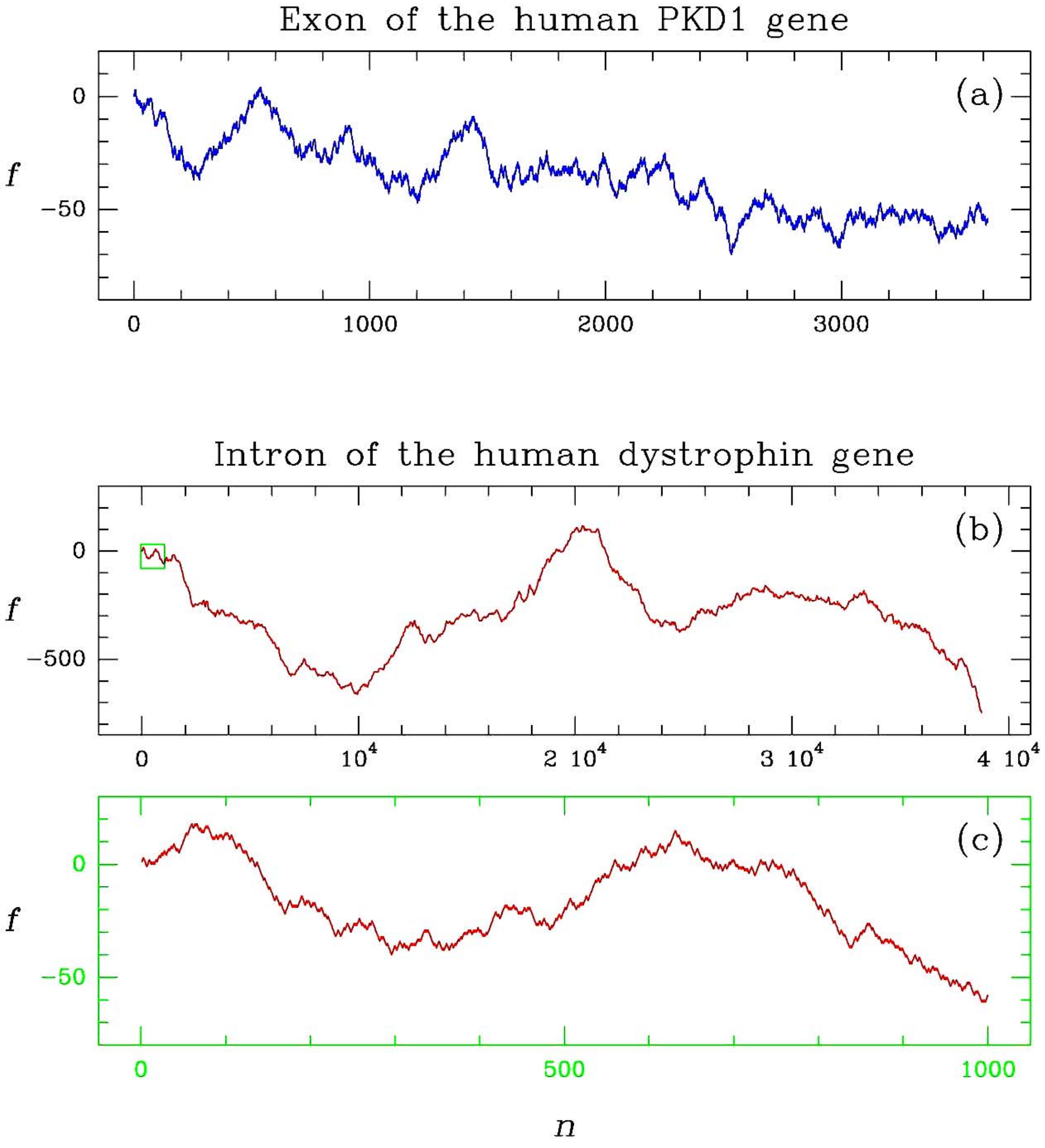




Exon of the human PKD1 gene

Intron of the human dystrophin gene

Most of the physicist works amount to characterizing the roughness of a DNA walk landscape

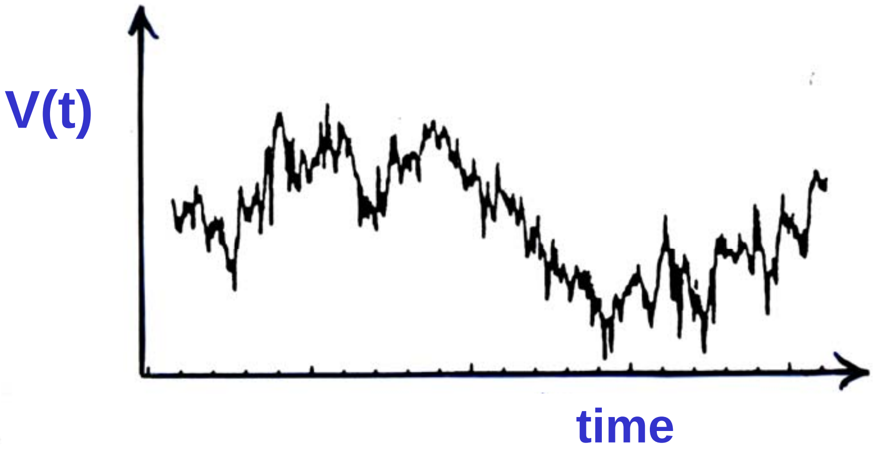


Exon of the human PKD1 gene

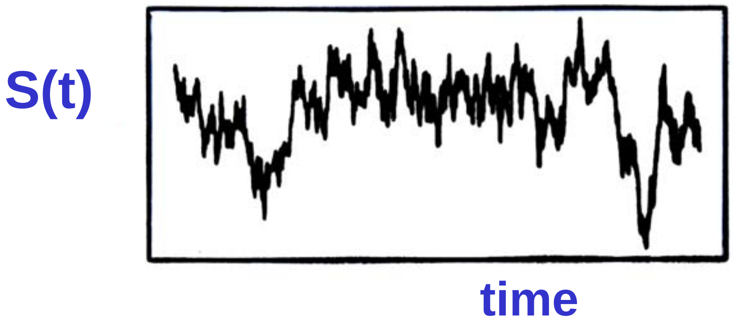
Intron of the human dystrophin gene

Most of the physicist works amount to characterizing the roughness of a DNA walk landscape

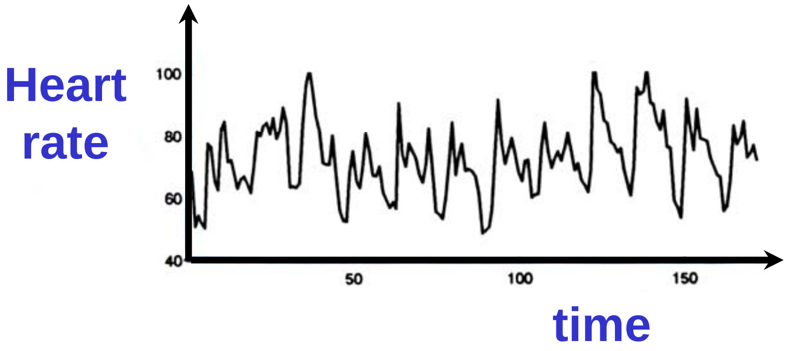
FRACTAL SIGNALS



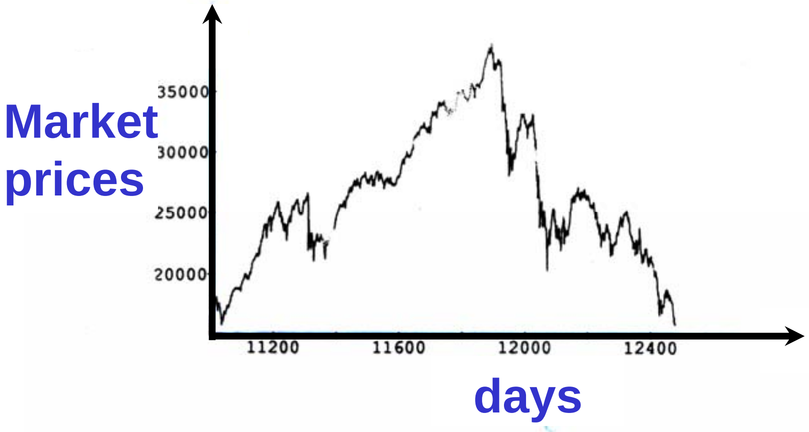
Turbulent velocity signal



Brownian signal
“ 1/f noise’’

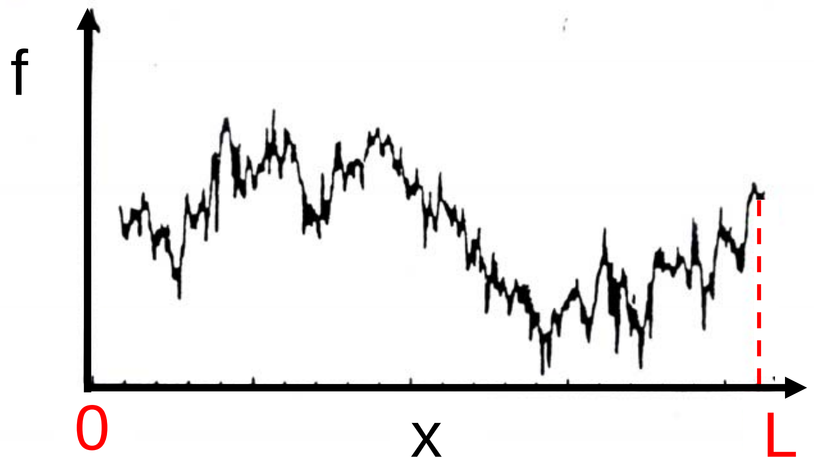


Medical signal



Financial time series

ROUGHNESS EXPONENT



- Root-mean square of the height fluctuations :

$$W(L) = \sqrt{< f^{2}(x) > - < f(x) >^{2}} \sim L^{H}$$

H = roughness exponent $\boxed{D_f = 2 - H}$

- Random walk
  - $0.5 < H < 1$ LONG RANGE CORRELATIONS (LRC)
  - $H = 0.5$ UNCORRELATED
  - $0 < H < 0.5$ ANTI-CORRELATIONS

- Power spectrum

$$S_f(k) \sim k^{-(2H+1)}$$

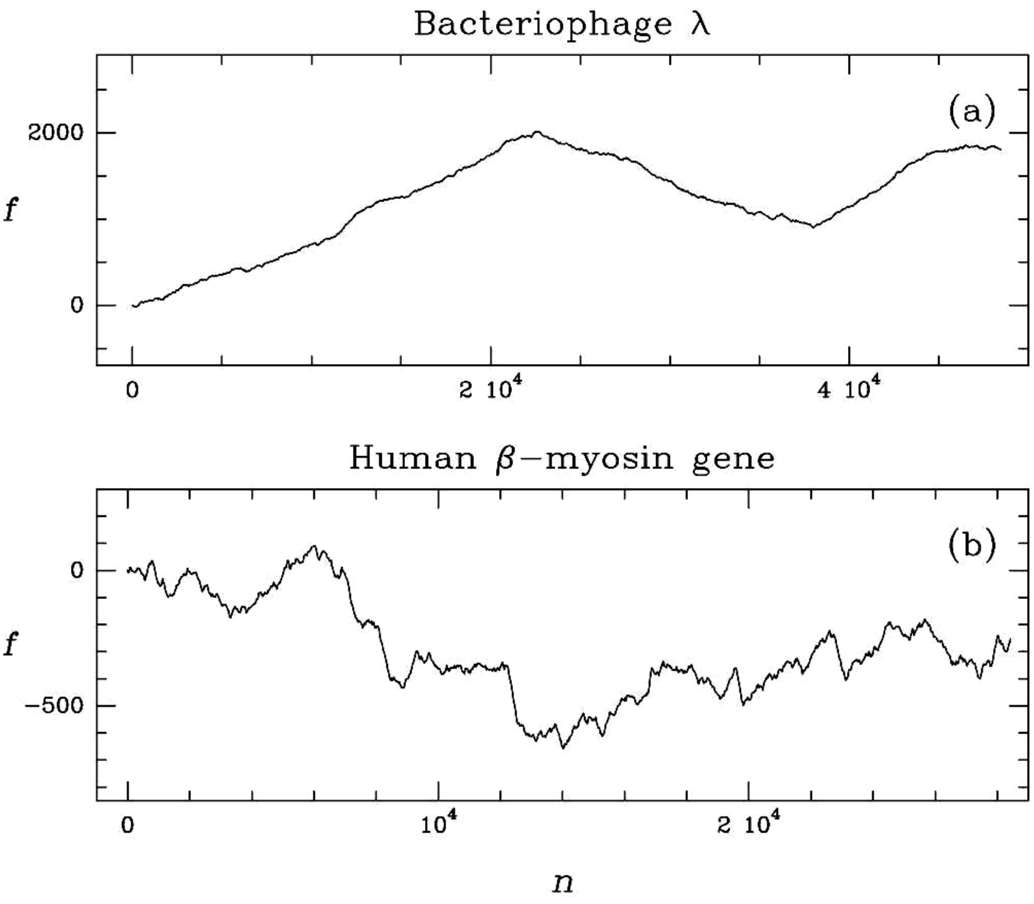
- Correlation function

$$C_f(l) = < f(x) f(x+l) > - < f(x) >^{2} \sim l^{2H}$$

Are the observed LRC a bias in the measurement ?

Is the mosaic structure of DNA enough to account for the observed misleading LRC in DNA sequences ?

Karlin and Brendel 93 :

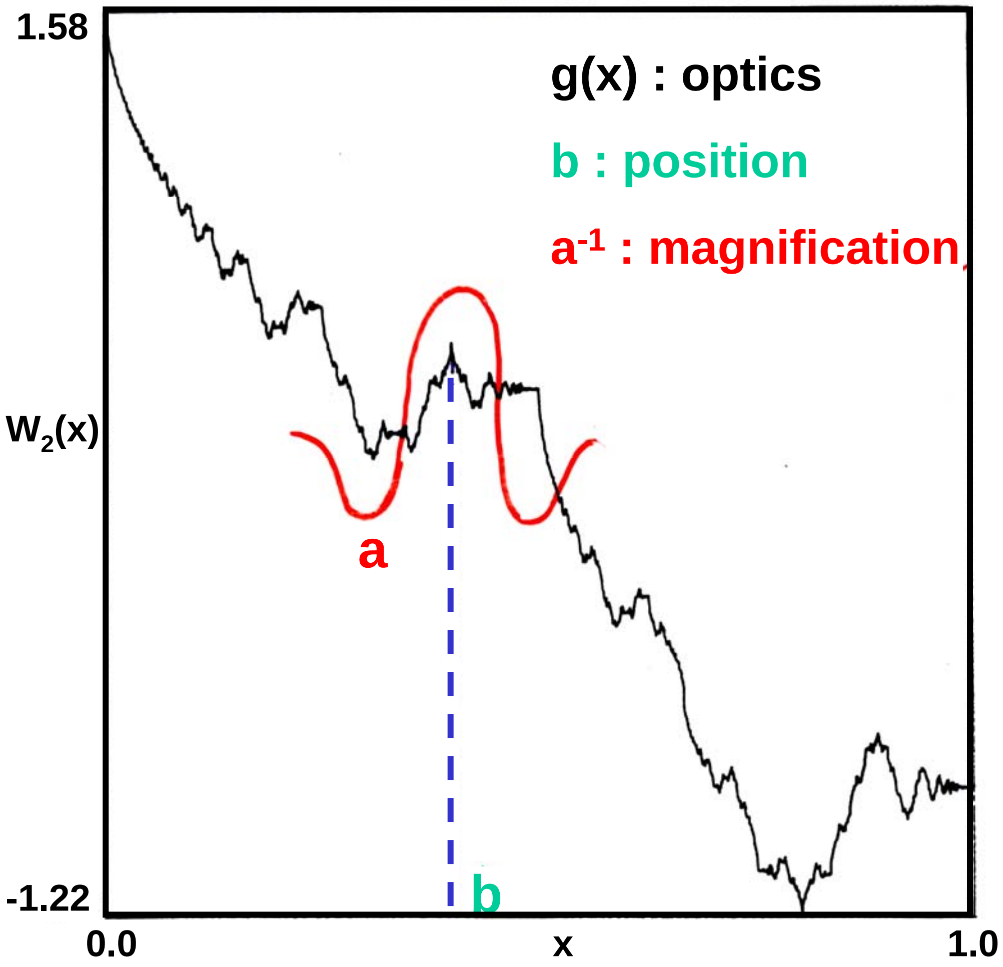


A specific analysing tool is needed to avoid confusing a biased uncorrelated random walk with an unbiased correlated random walk

WAVELET ANALYSIS OF FRACTAL SIGNALS

$$T_g\,(a,b) = \frac{1}{a}\int g^*\left(\frac{x-b}{a}\right) f(x)\,dx$$

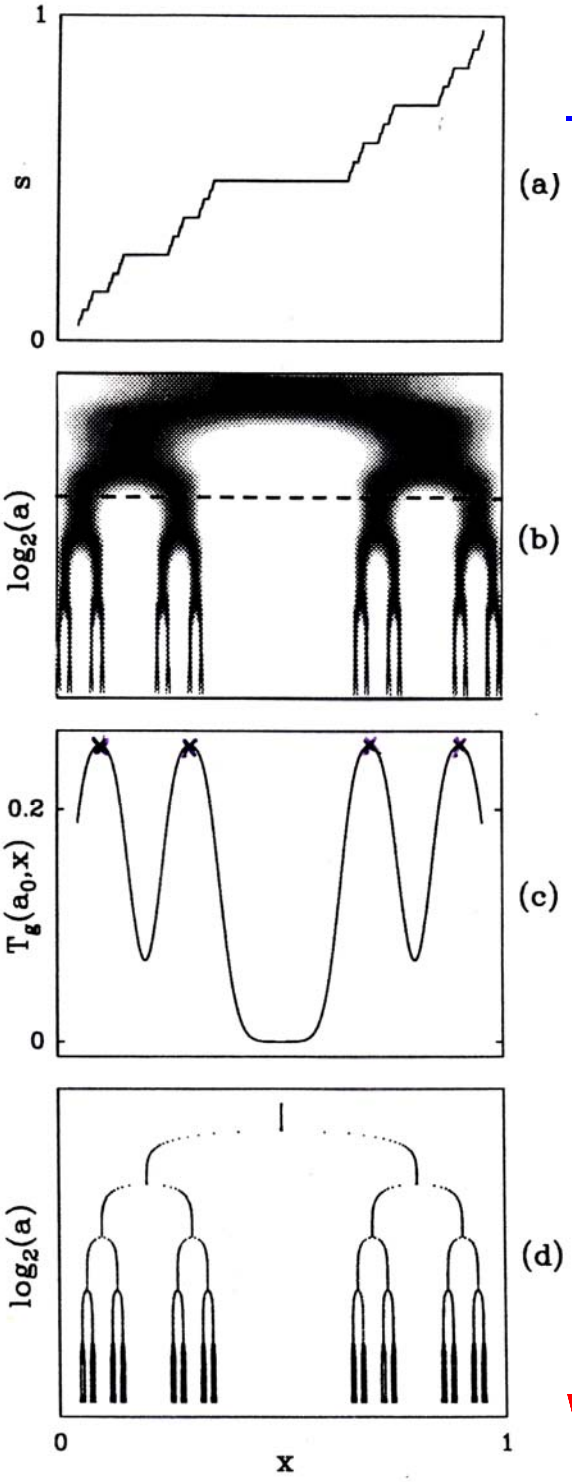
Mathematical microscope



“ Singularity scanner’’

The wavelet transform allows us to **LOCATE** (**b**) the singularities of f and to **ESTIMATE** (**a**) their strength h(x) (Hölder exponent)

CONTINUOUS WAVELET TRANSFORM OF THE TRIADIC DEVIL'S STAIRCASE



THE DEVIL'S STAIRCASE

$$F(x) = \int_{-\infty}^{x} d\mu(x)$$

WAVELET TRANSFORM REPRESENTATION

WAVELET TRANSFORM MODULUS MAXIMA (WTMM)

WTMM SKELETON

|||

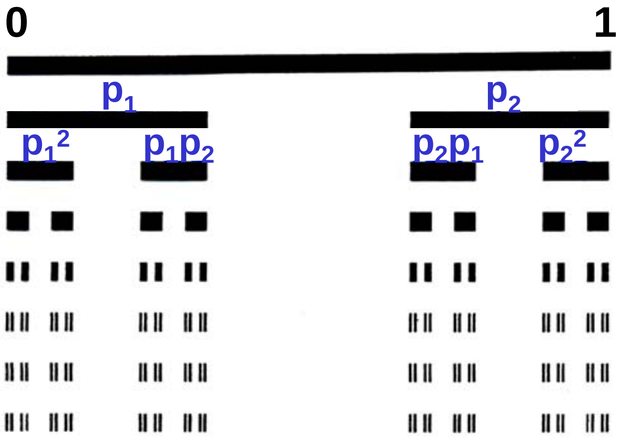
WTMM SKELETON OF THE TRIADIC CANTOR SET

F(x) is continuous but non differentiable. F'(x)=0 almost everywhere. Its continuous variation occurs over a set of Lebesgue measure = 0 and dimension $D_F = \log 2 / \log 3$

Fractal measures

- Invariant measures associated with the strange attractors of discrete dynamical systems
- Turbulent energy dissipation

TRIADIC CANTOR SET



UNIFORM

$p_1 = p_2 = ½$

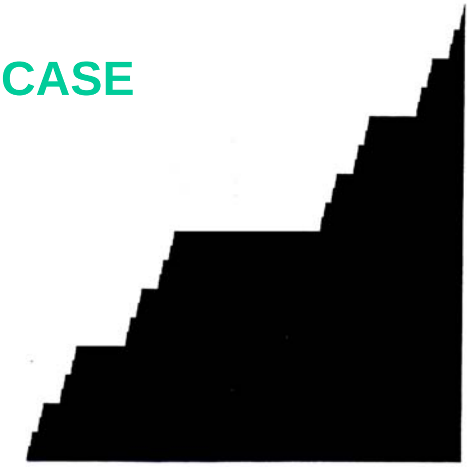
MULTIFRACTAL

$p_1 \neq p_2$

Fractal signals

- Weierstrass functions
- Fractional Brownian motions
- Turbulent signals

DEVIL’S STAIRCASE

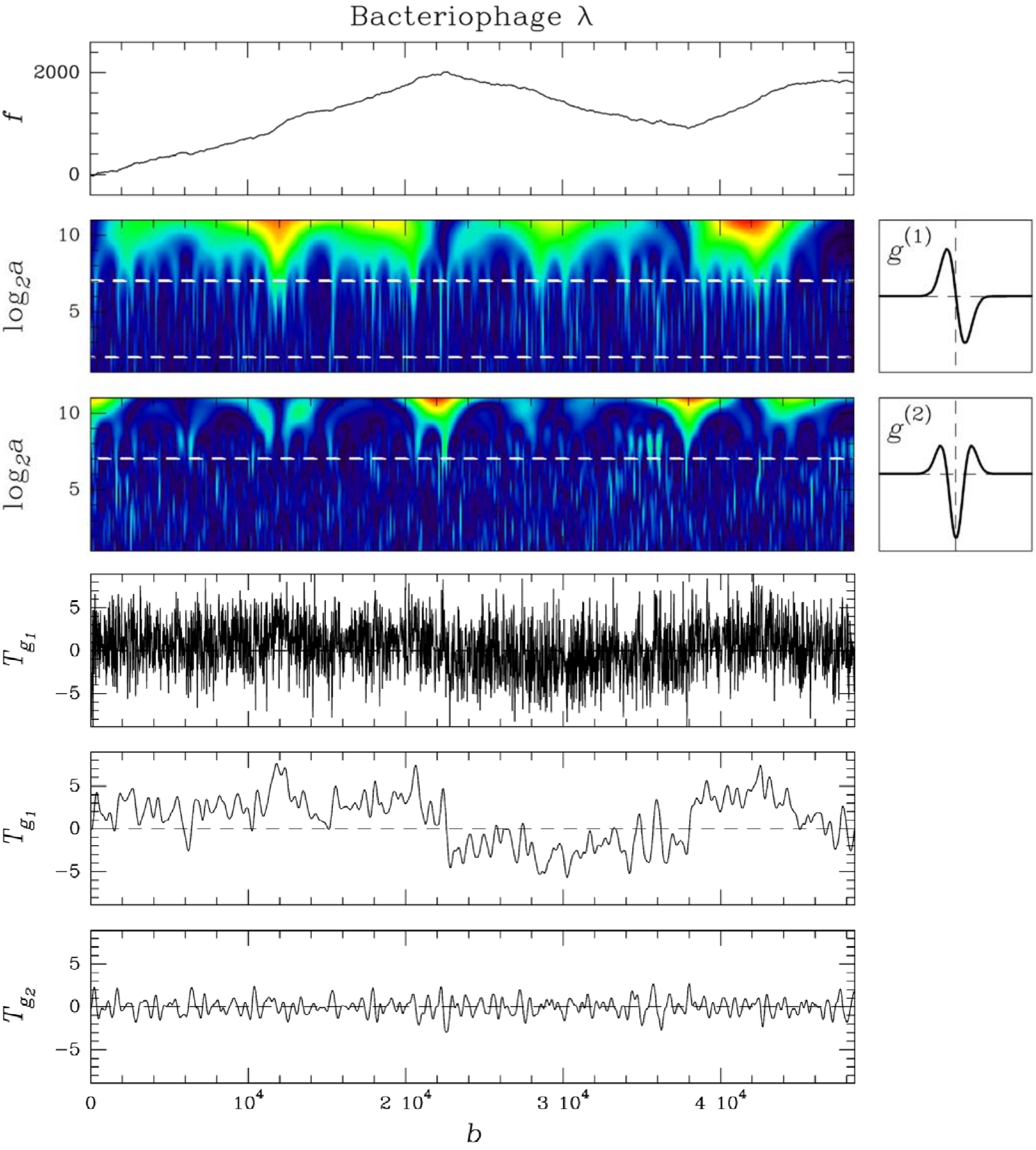


$$F(x) = \int_{-\infty}^{x} d\mu(x)$$

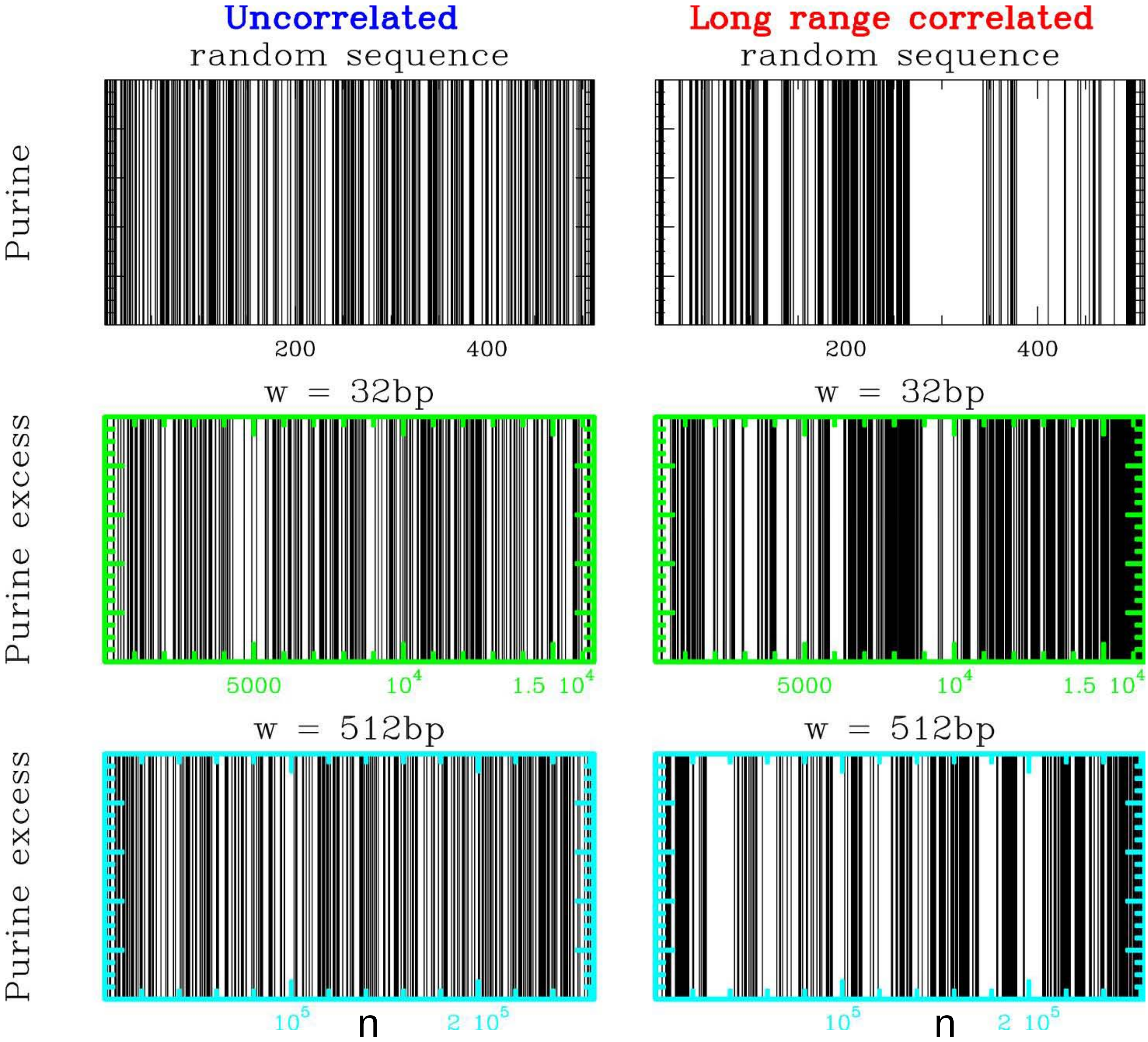
Characteristic function of μ

F(x) is continuous but non differentiable. F’(x)=0 almost everywhere. Its continuous variation occurs over a set of Lebesgue measure = 0 and dimension D_F = log 2 / log 3

Wavelet analysis of the DNA sequence of the bacteriophage λ

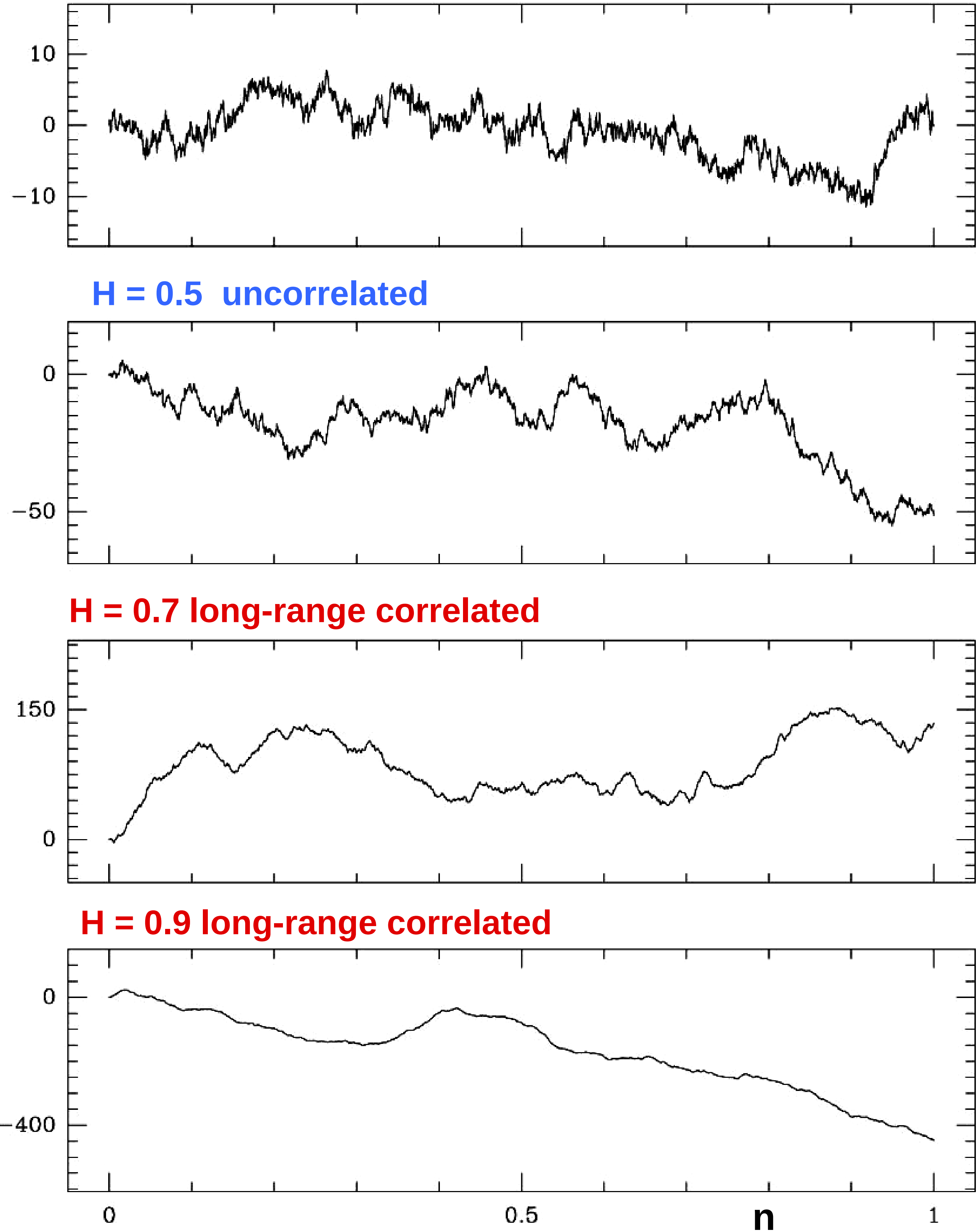


SYNTHETIC DNA SEQUENCES



SYNTHETIC DNA WALKS

Fractional Brownian motions : B_H



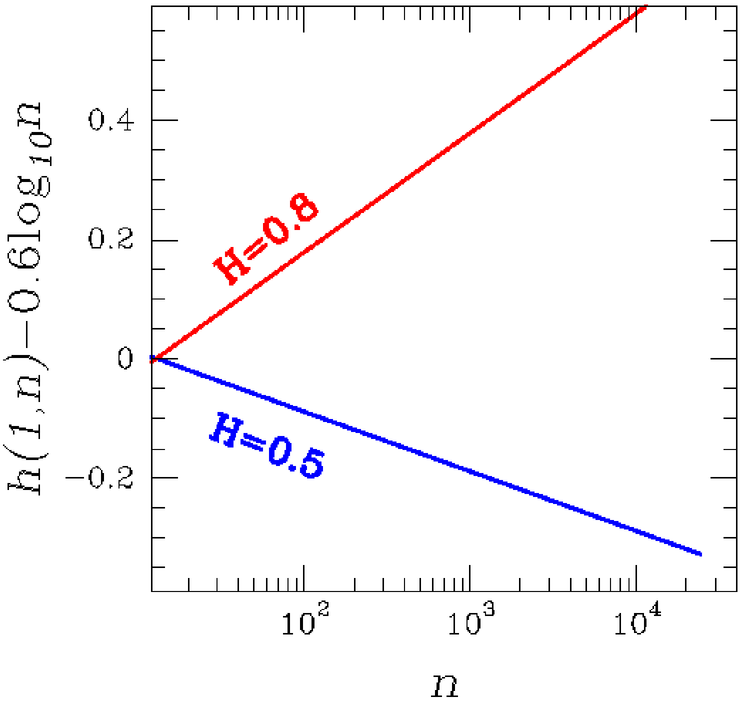
H = 0.3 anti-correlated

H = 0.5 uncorrelated

H = 0.7 long-range correlated

H = 0.9 long-range correlated

A UNIQUE WAY TO DISPLAY RESULTS

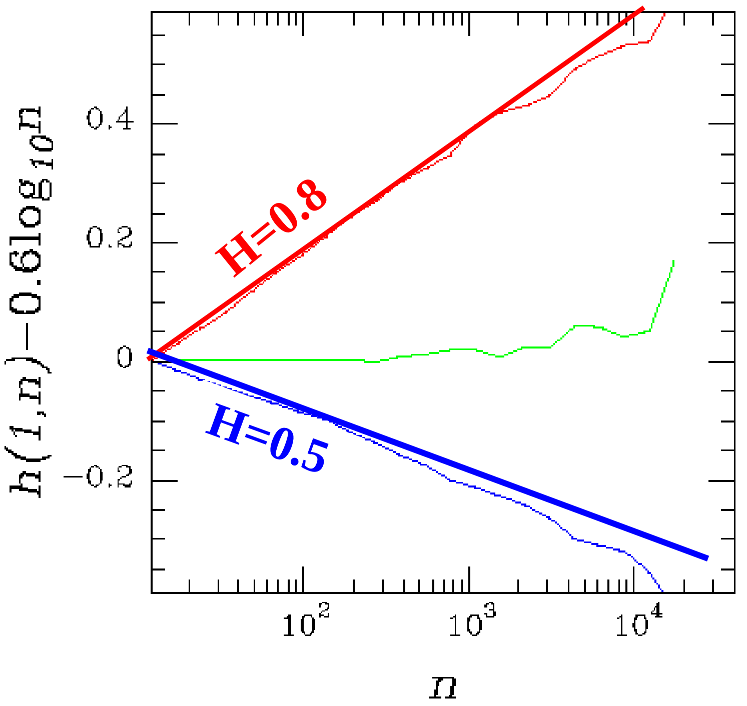


1. Straight line ⇔ scale invariance properties

2. The slope of a linear behavior gives the roughness exponent H

$$\begin{cases} H = 0.5 & \text{No LRC} \\ H > 0.5 & \text{LRC} \end{cases}$$

A UNIQUE WAY TO DISPLAY RESULTS



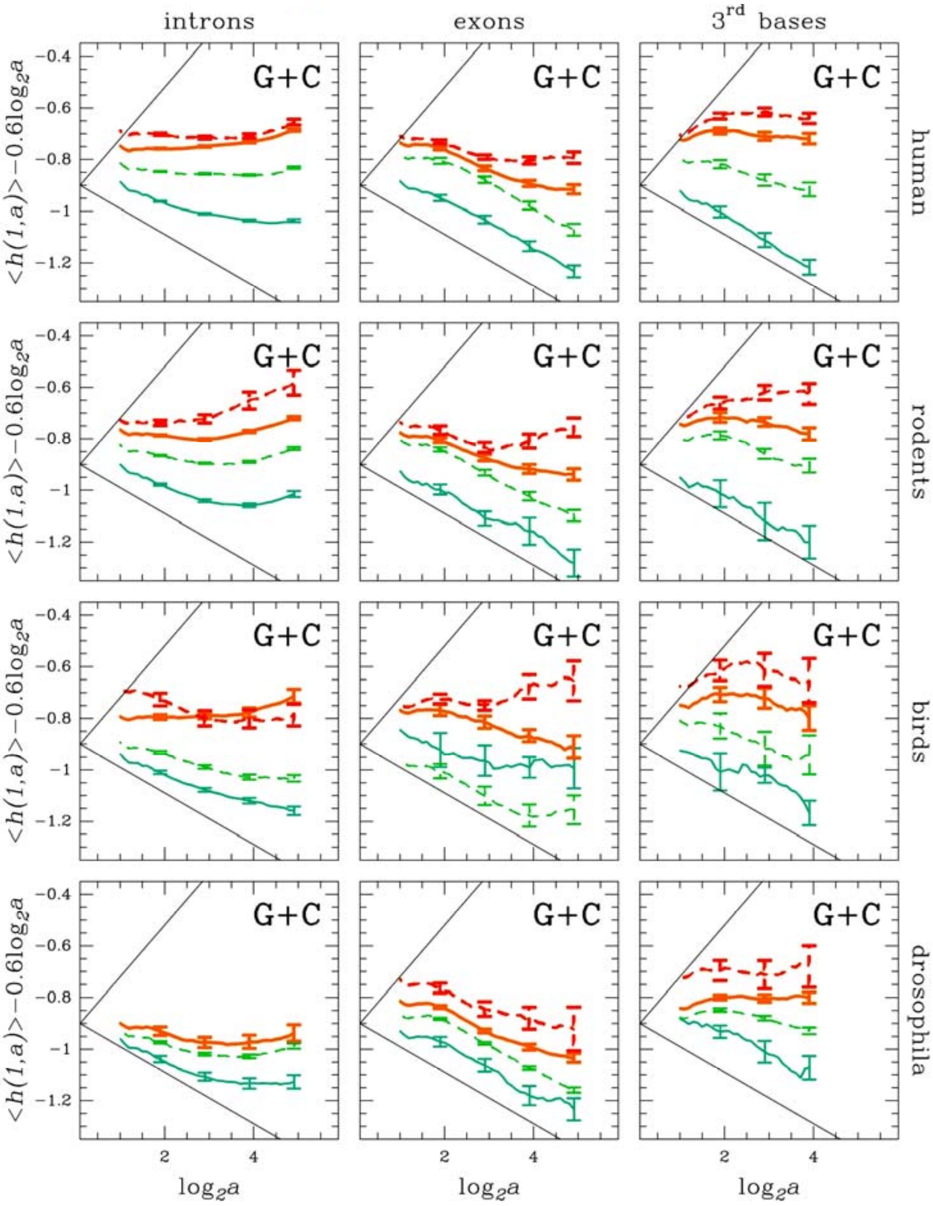
1. Straight line ⇔ scale invariance properties

2. The slope of a linear behavior gives the roughness exponent H

$$\begin{cases} H = 0.5 & \text{No LRC} \\ H > 0.5 & \text{LRC} \end{cases}$$

LRC AND THE ISOCHORE STRUCTURE OF WARM BLOODED VERTEBRATES

G + C poor ***G + C*** rich



LRC increase with the $G + C$ content of isochores

This result remains valid for genomes that don't possess an isochore structure !

Which biological mecanisms can account for LRC in DNA sequences

- Genomes dynamics and plasticity

  Point mutation

  Insertion, deletion

  Transposition

  Duplication of exons, genes or chromosomes

  Recombinaison

  Generalized Lévy walk model (Buldyrev *et al.* 93)

  Length distribution of protein coding segments (Herzel and Große 97)

- Compaction constraints - Accession to information

  Nucleosome

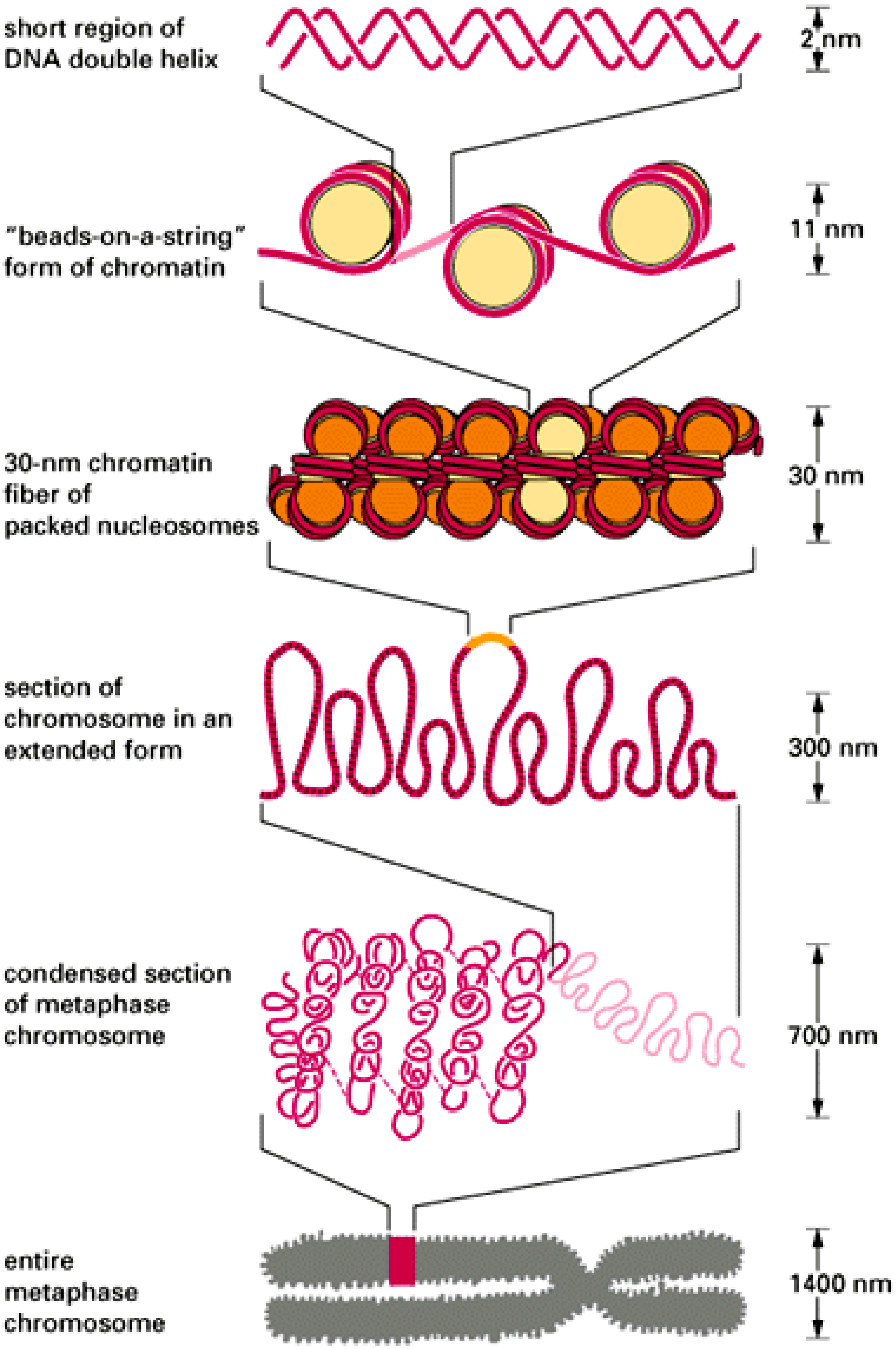
  Chromatine fiber

  Higher order folding up to the metaphase chromosome

  Fractal model of chromosomes (Takahashi 89)

  Crumpled globule model (Grosberg *et al.* 93)

HIERARCHICAL STRUCTURE OF EUCARYOTIC DNA



short region of DNA double helix

2 nm

"beads-on-a-string" form of chromatin

11 nm

30-nm chromatin fiber of packed nucleosomes

30 nm

section of chromosome in an extended form

300 nm

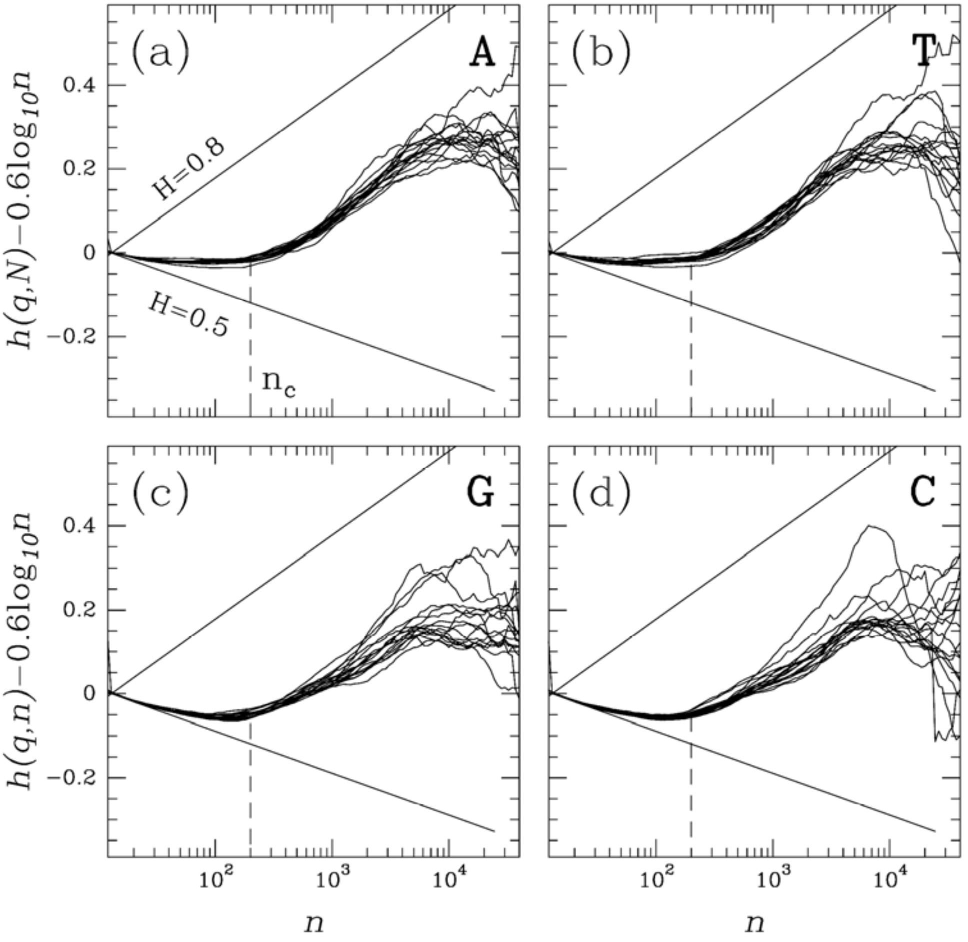
condensed section of metaphase chromosome

700 nm

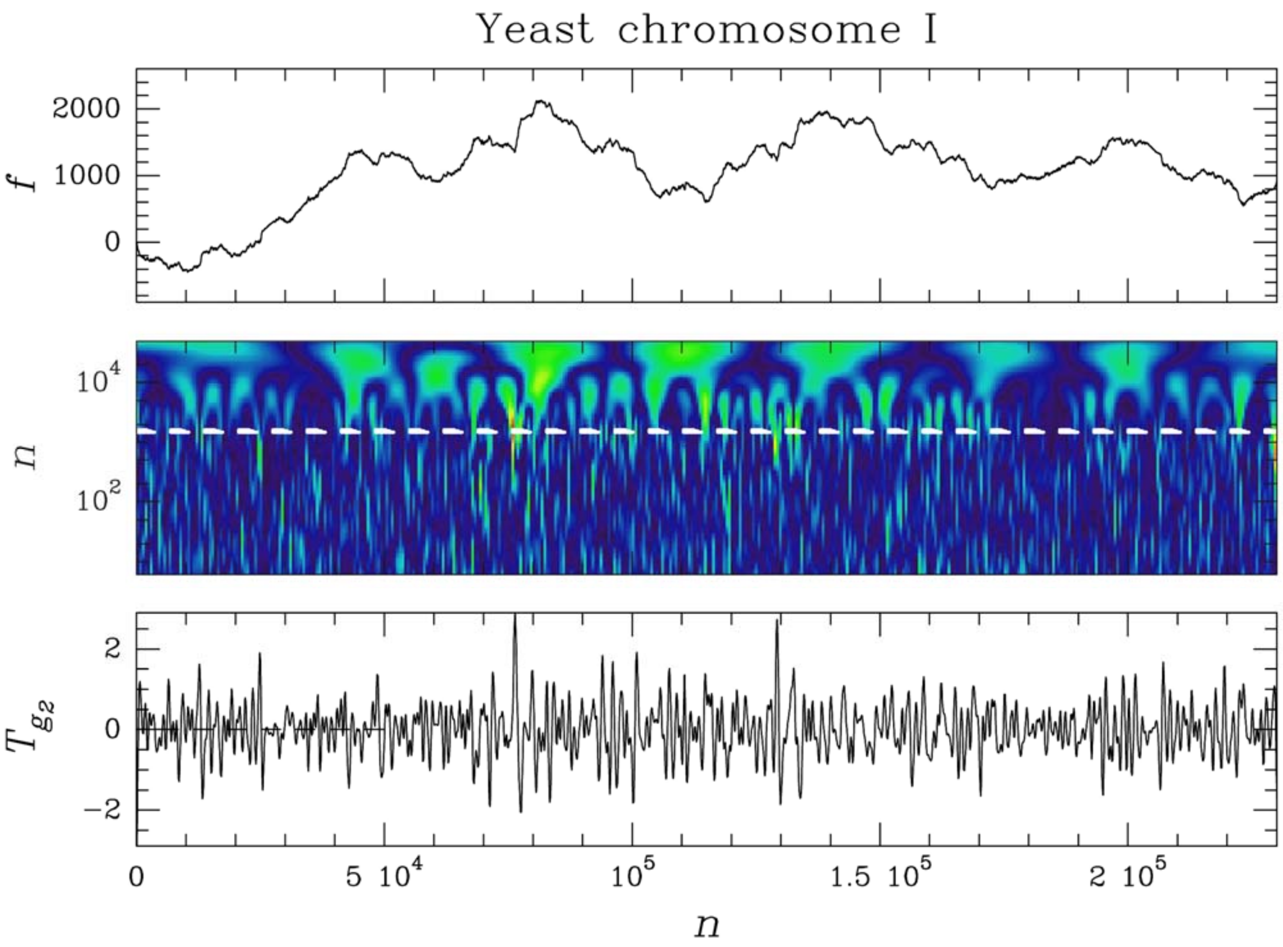
entire metaphase chromosome

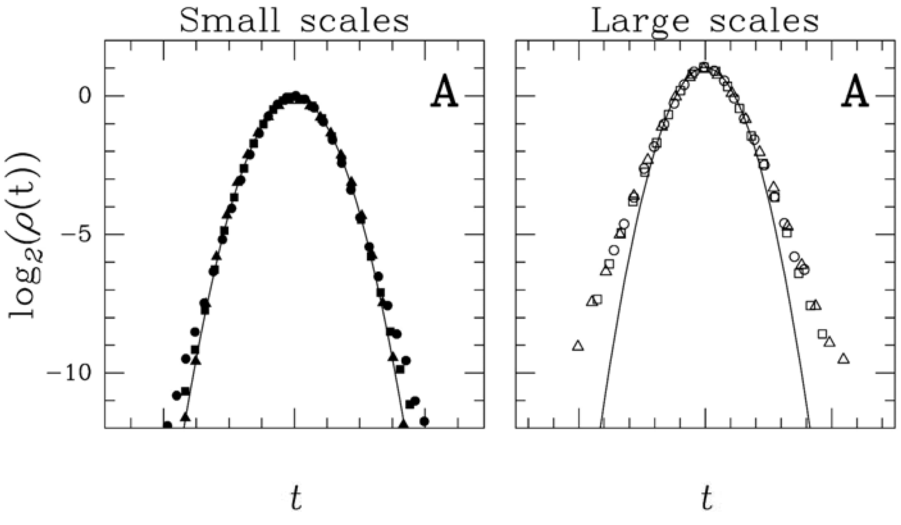
1400 nm

Statistical analysis of the eukaryotic genome of *Saccharomyces cerevisiae*



Universality between the 16 chromosomes of yeast
Universality between the 4 mononucleotidic codings
$n_c \sim 200$bp is a characteristic length scale

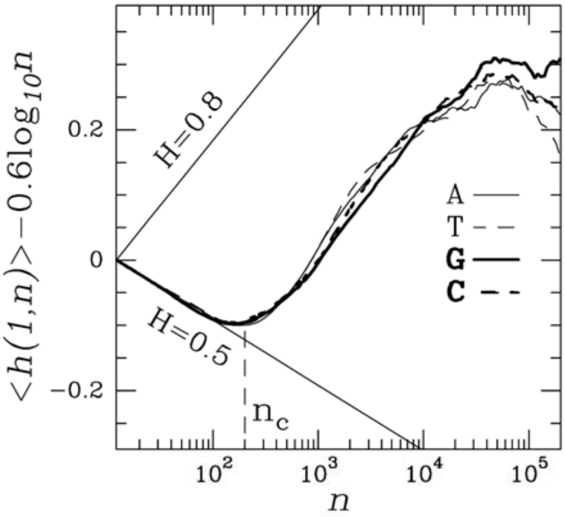




Gaussian statisics at small scales ($n \leq$ 200bp)

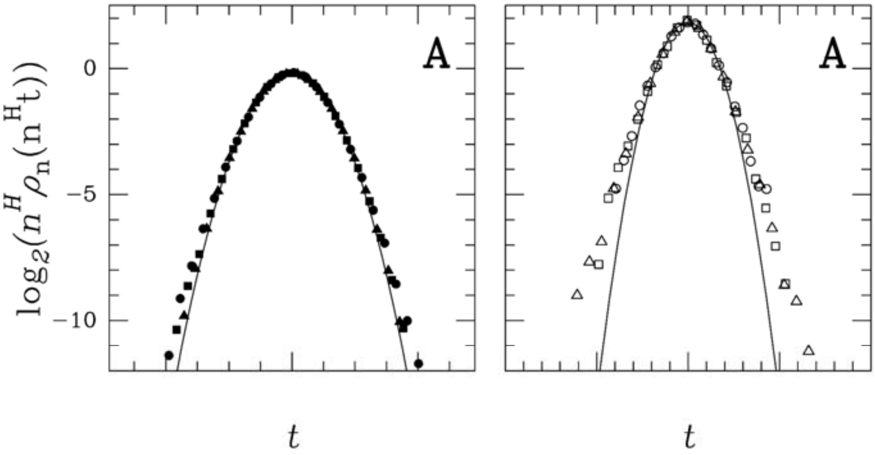
Non Gaussian (fat tails) statistics at large scale ($n \geq$ 200bp)

STATISTICAL ANALYSIS OF THE BACTERIAL GENOME OF *Escherichia coli*



Universality between the 4 mononucleotidic codings and with the eukaryotic genome of yeast

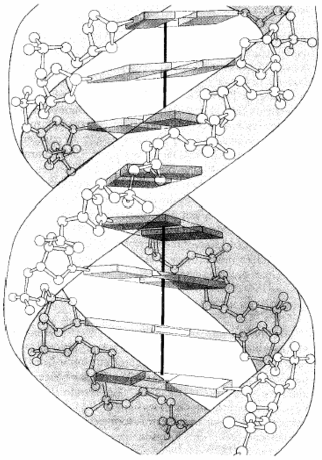
$n_c \sim 200$bp is a characteristic length scale

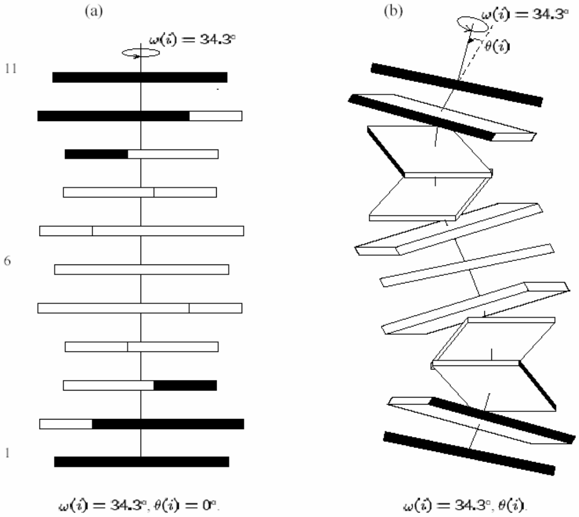


Gaussian statisics at small scales ($n \leq 200$bp): $H = 0.5$

Non Gaussian (fat tails) statistics at large scale ($n \geq 200$bp): $H = 0.75$

DNA walks that reflect the structure of the DNA polymer



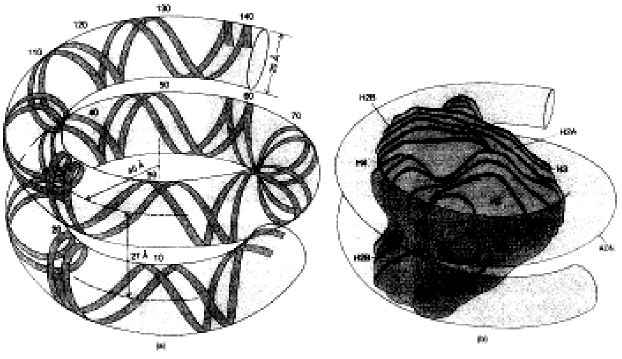

(a)
$\omega(i) = 34.3°$
11
6
1
$\omega(i) = 34.3°, \theta(i) = 0°$.
(b)
$\omega(i) = 34.3°$
$\theta(i)$
$\omega(i) = 34.3°, \theta(i)$.

2 trinucleotide codings based on experiments :

Trinucleotide		
	PNuc	DNase I
AAA/TTT	0.0	0.1
AAC/GTT	3.7	1.6
AAG/CTT	5.2	4.2
AAT/ATT	0.7	0.0
ACA/TGT	5.2	5.8
ACC/GGT	5.4	5.2
ACG/CGT	5.4	5.2
ACT/AGT	5.8	2.0
AGA/TCT	3.3	6.5
AGC/GCT	7.5	6.3
AGG/CCT	5.4	4.7
ATA/TAT	2.8	9.7
ATC/GAT	5.3	3.6
ATG/CAT	6.7	8.7
CAA/TTG	3.3	6.2
CAC/GTG	6.5	6.8

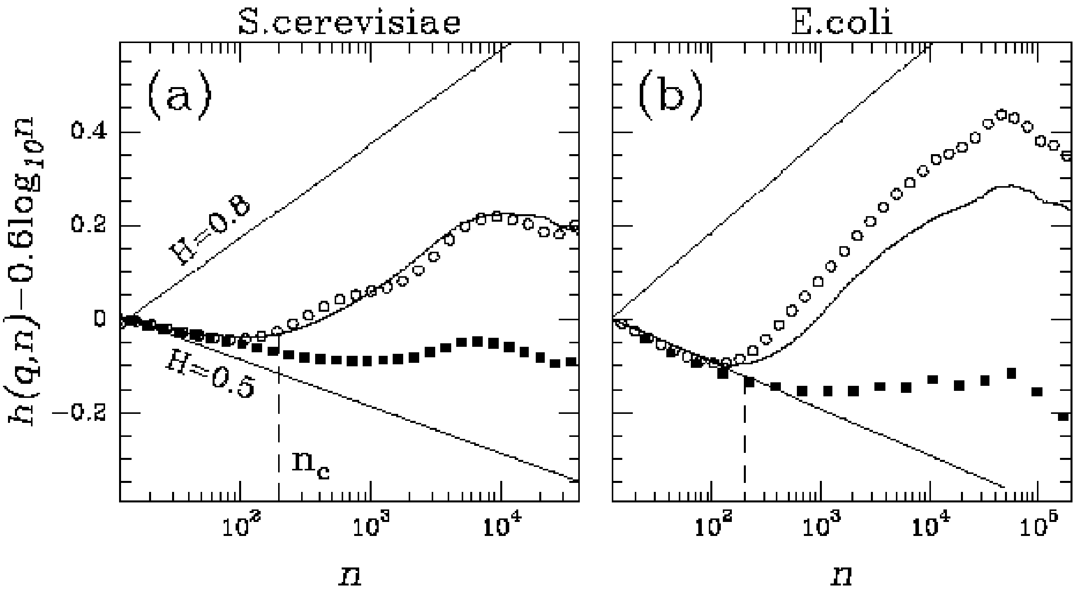
Trinucleotide		
	PNuc	DNase I
CAG/CTG	4.2	9.6
CCA/TGG	5.4	0.7
CCC/GGG	6.0	5.7
CCG/CGG	4.7	3.0
CGA/TCG	8.3	5.8
CGC/GCG	7.5	4.3
CTA/TAG	2.2	7.8
CTC/GAG	5.4	6.6
GAA/TTC	3.0	5.1
GAC/GTC	5.4	5.6
GCA/TGC	6.0	7.5
GCC/GGC	10.0	8.2
GGA/TCC	3.8	6.2
GTA/TAC	3.7	6.4
TAA/TTA	2.0	7.3
TCA/TGA	5.4	10.0

1. Nucleosome positioning model (PNuc)



related to curvature ?

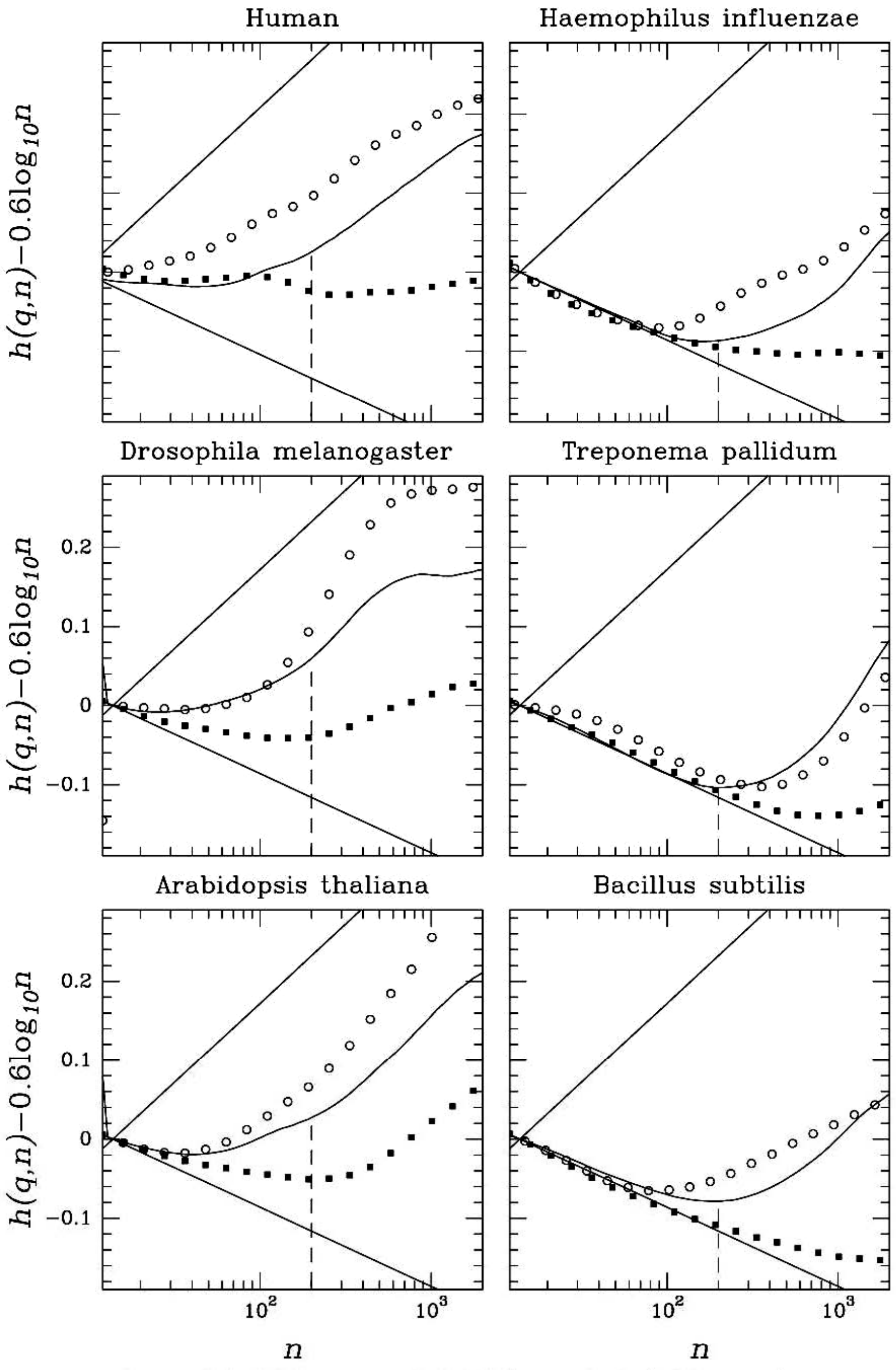
2. DNase I digestion data
related to bending propensity



(——) DNA text, (○) PNuc, (■) DNase I

Hypothesis: LRC in the small scales regime is the signature of of the nucleosomal structure

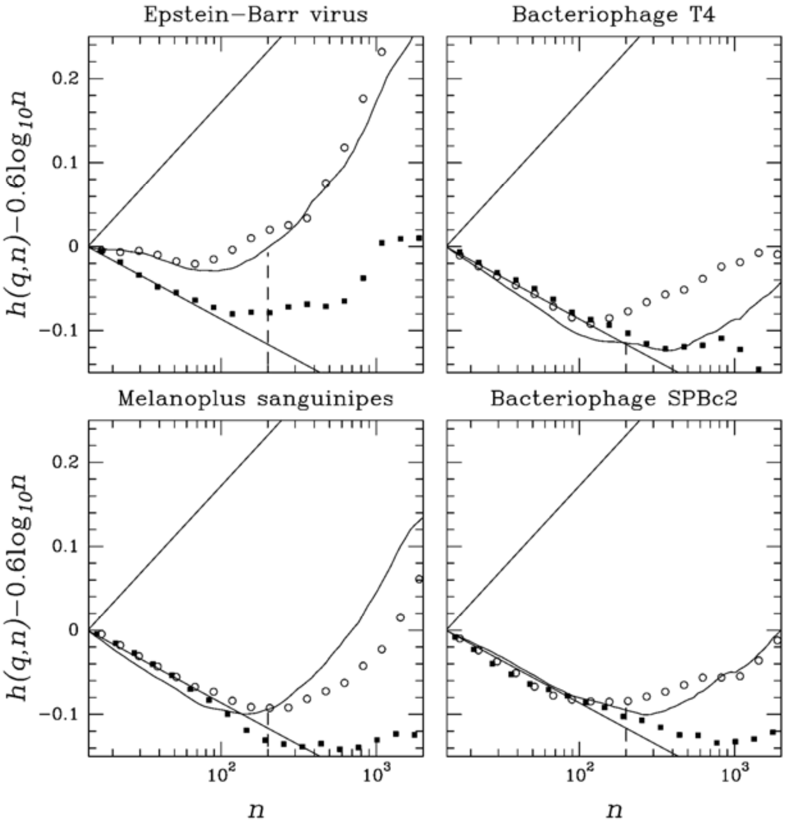
Eucaryotes Bacteria



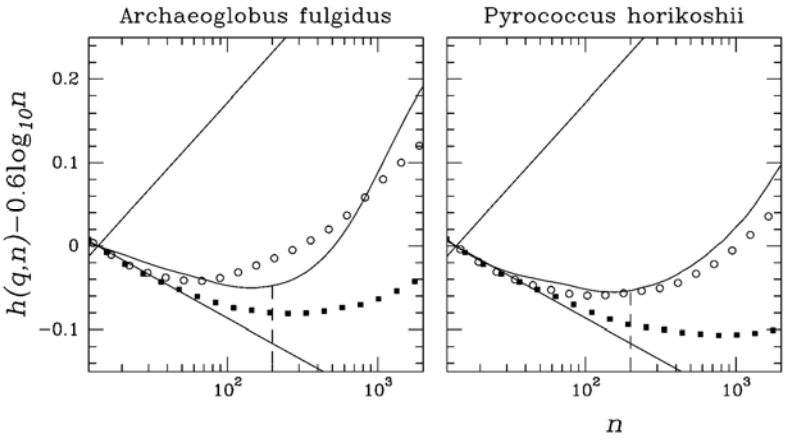
(——) DNA text, (○) PNuc, (■) DNase I

Nucleosomes No nucleosomes

Small scales LRC are related to nucleosome like structures



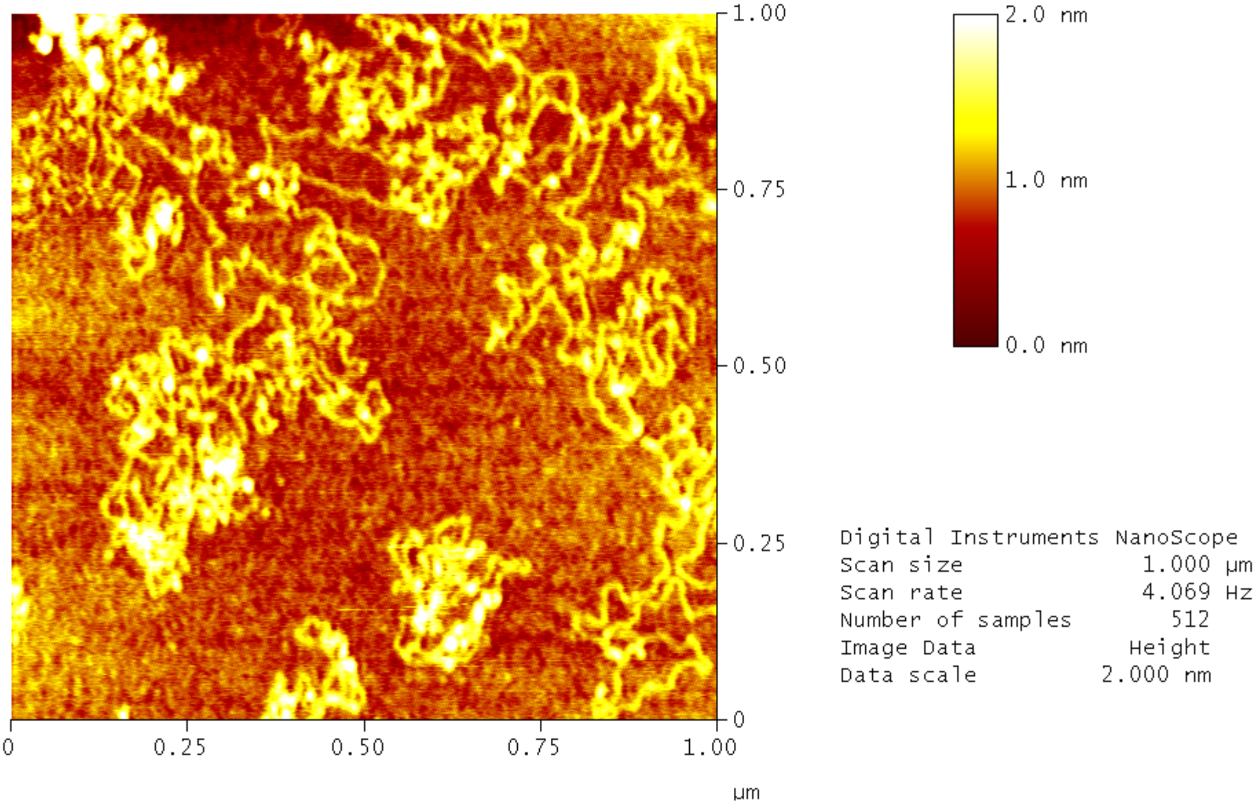
Pox virus don’t display LRC in the small scale regime



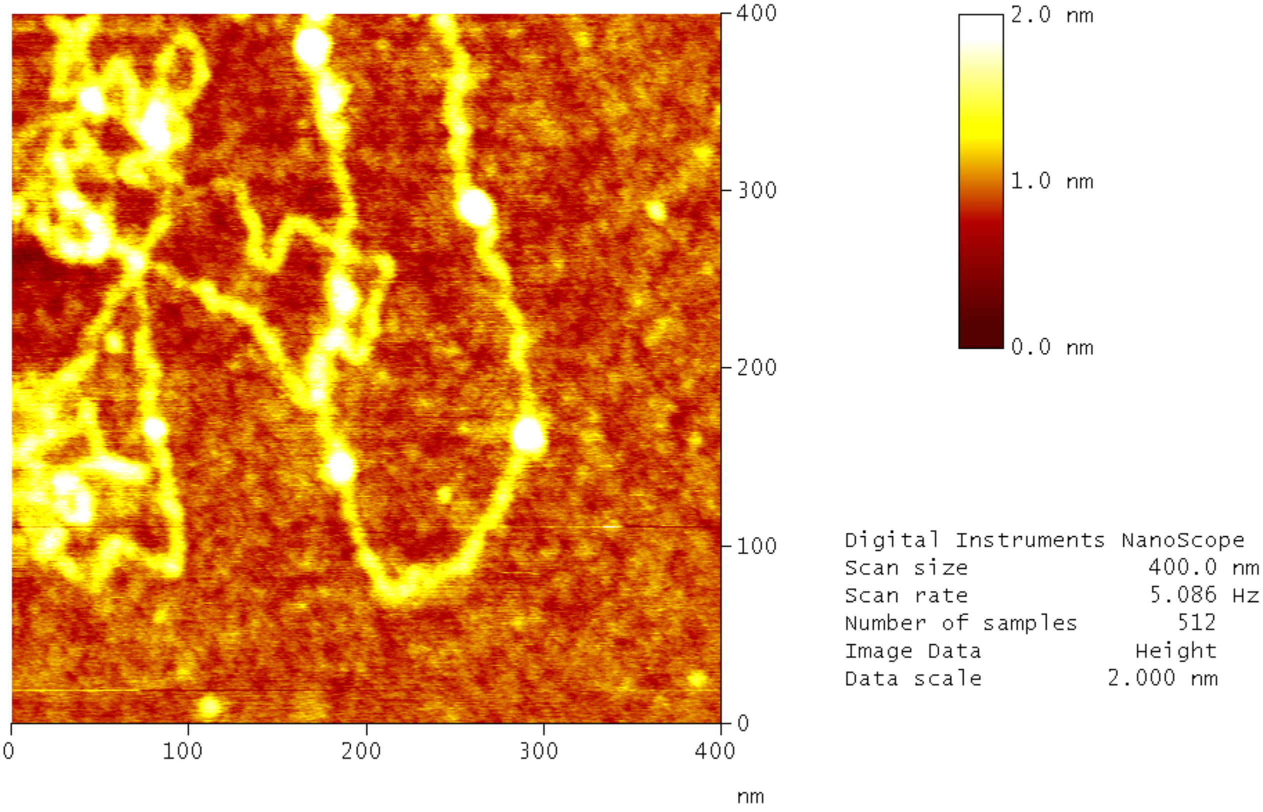
Archaebacteria display LRC in the small scale regime

AFM visualisation of a reconstituted chromatin fiber

Pierre-Louis Porté, Emeline Fontaine, Cendrine Moskalenko



```
Digital Instruments NanoScope
Scan size              1.000 µm
Scan rate              4.069 Hz
Number of samples        512
Image Data             Height
Data scale           2.000 nm
```



```
Digital Instruments NanoScope
Scan size              400.0 nm
Scan rate              5.086 Hz
Number of samples        512
Image Data             Height
Data scale           2.000 nm
```

Images obtained in ‘Tapping Mode’ in air

Linear DNA (2500 bp) positioning nucleosomes

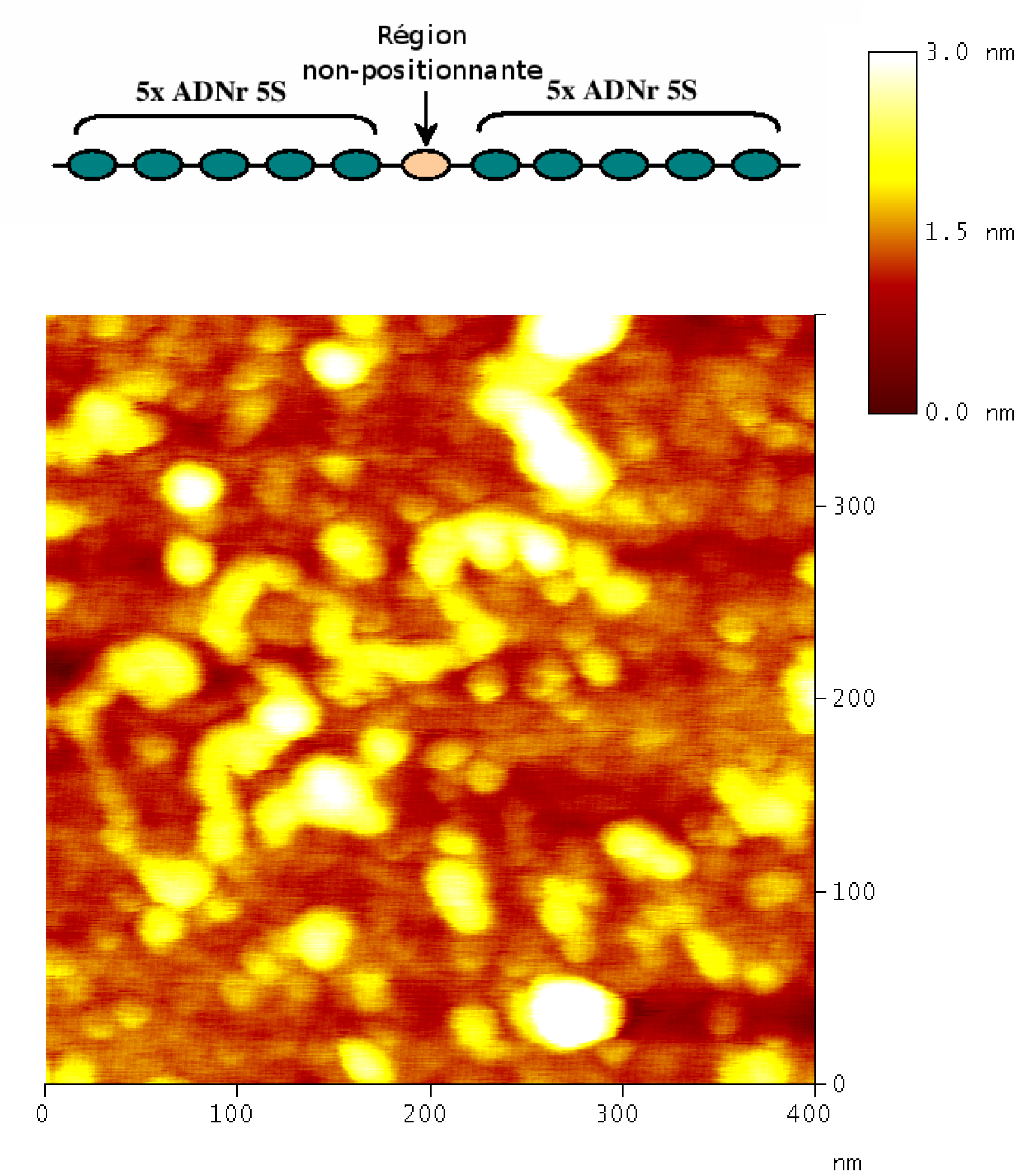


Image obtained in 'Tapping Mode' in air

Linear DNA (2500 bp) positioning nucleosomes

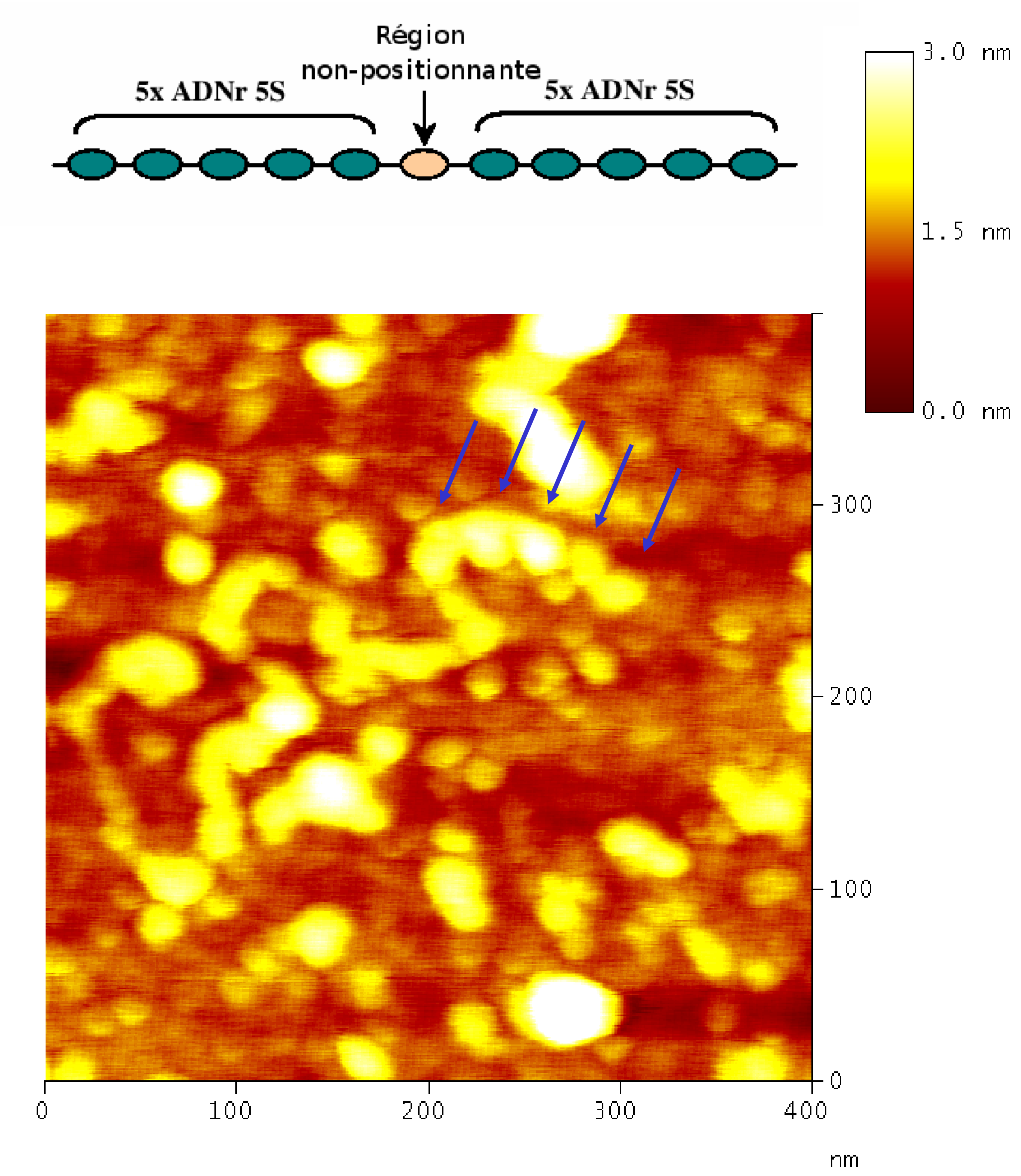
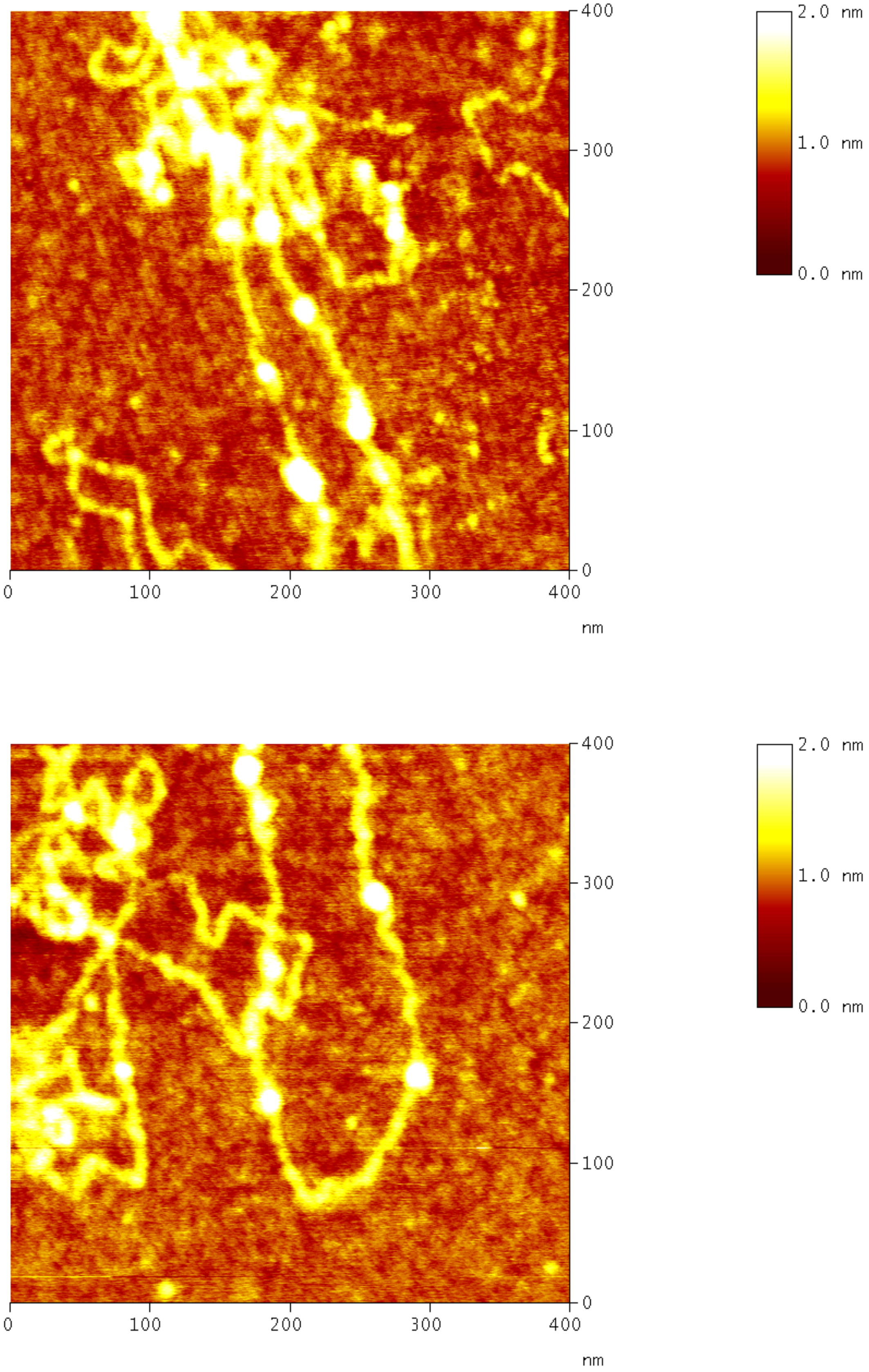


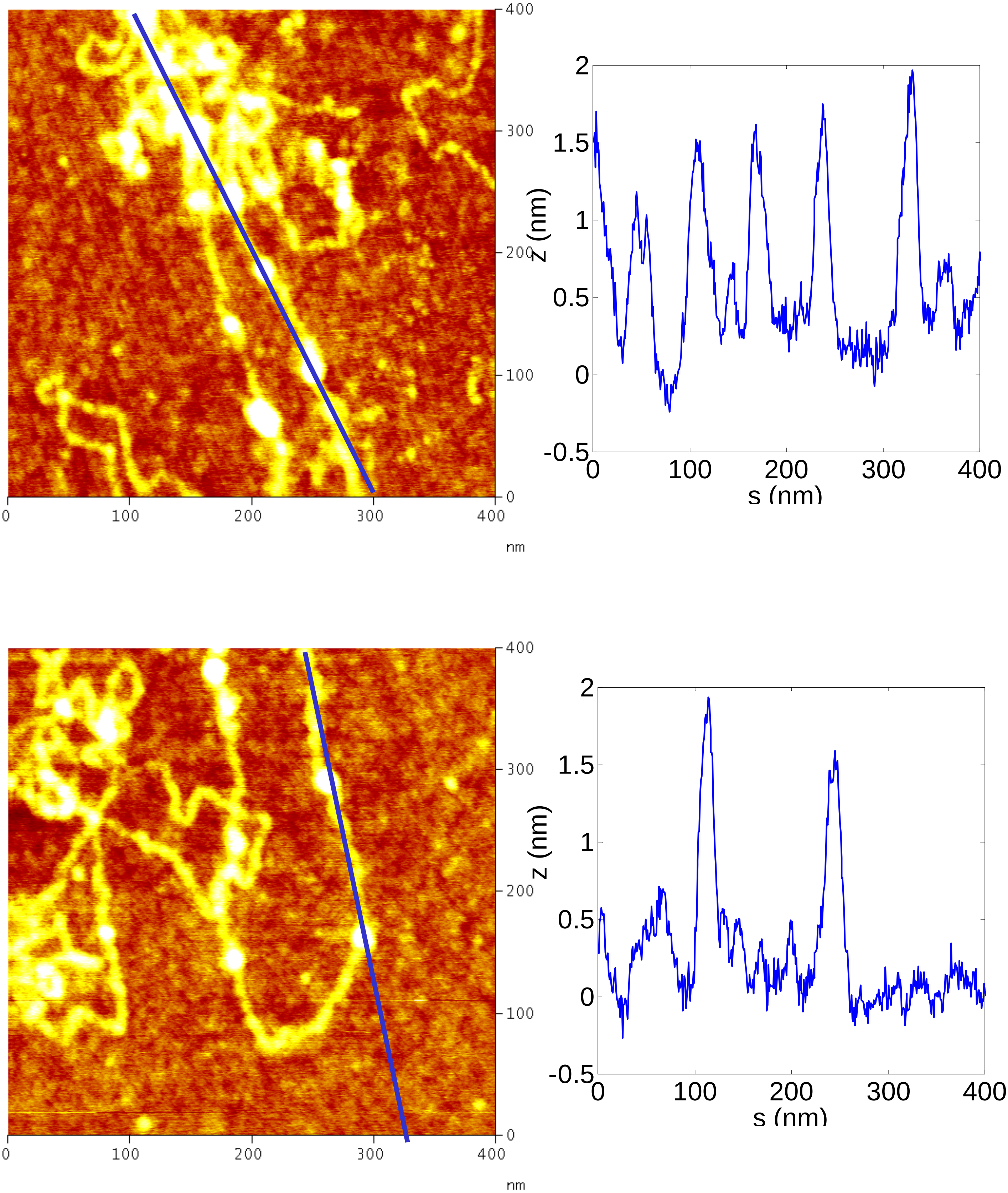
Image obtained in ‘Tapping Mode’ in air

Plasmid DNA (3200 bp) + nucleosomes



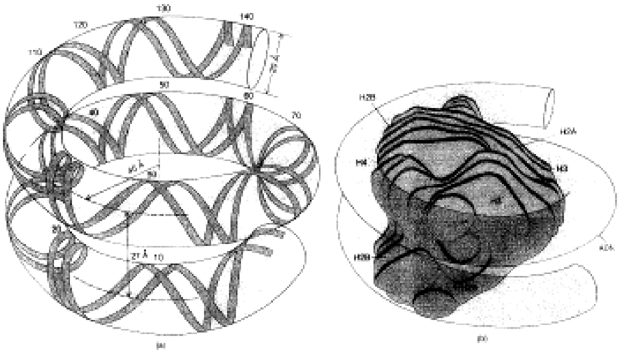
Images obtained in ‘Tapping Mode’ in air

Plasmid DNA (3200 bp) + nucleosomes



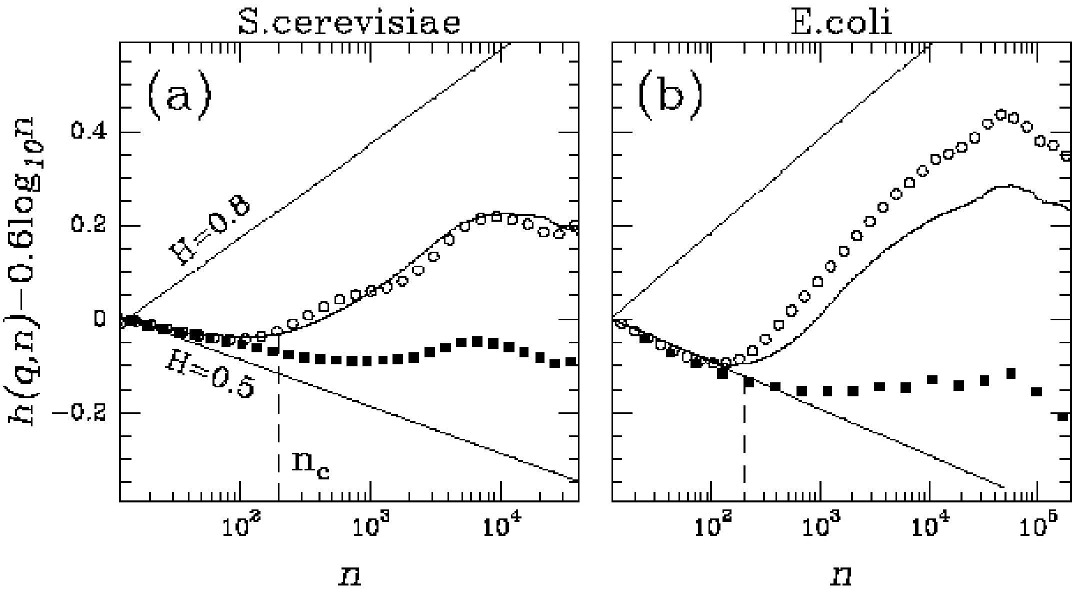
Images obtained in 'Tapping Mode' in air

1. Nucleosome positioning model (PNuc)



related to curvature ?

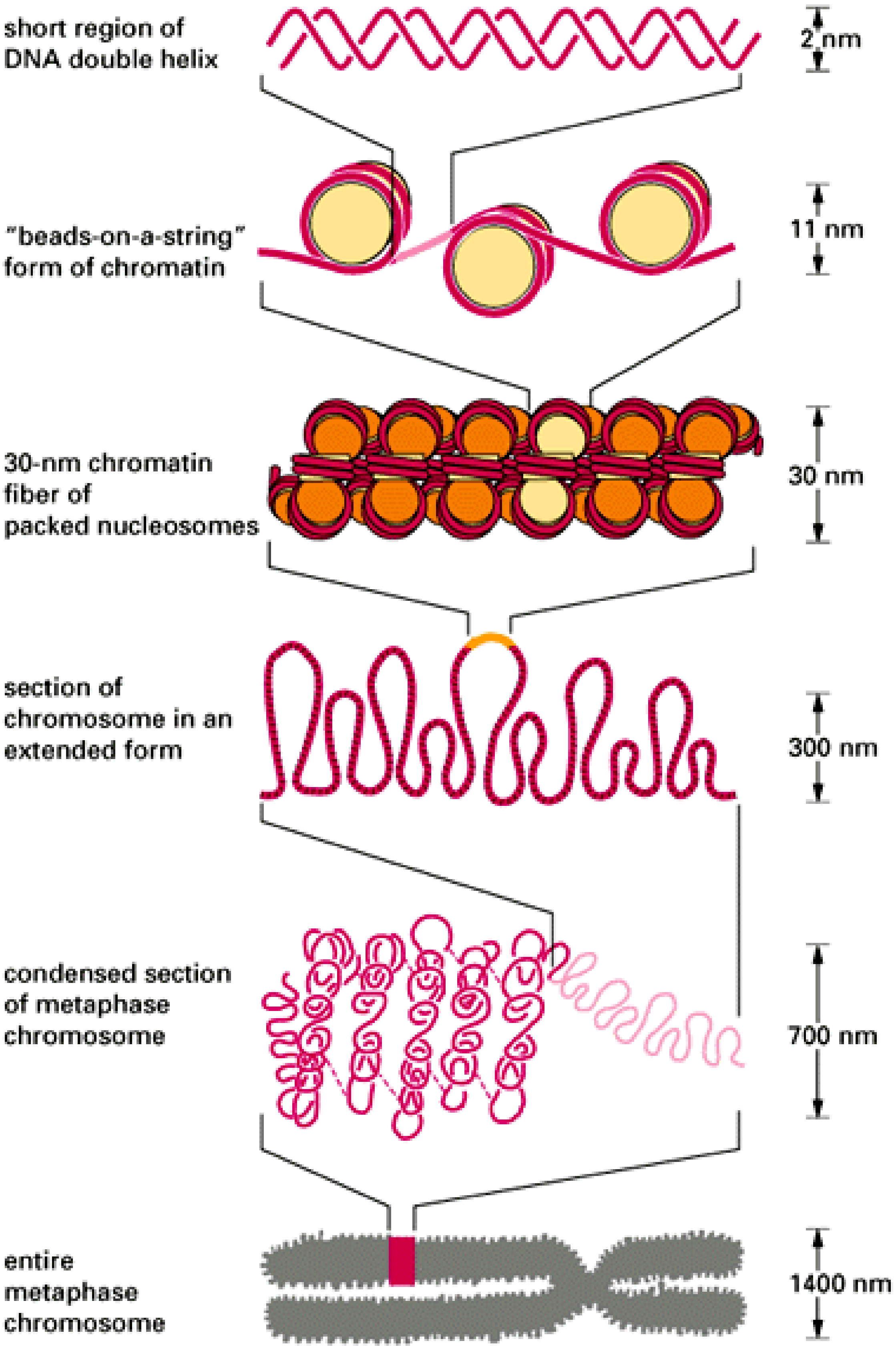
2. DNase I digestion data
related to bending propensity



(——) DNA text, (○) PNuc, (■) DNase I

Hypothesis: LRC in the small scales regime is the signature of of the nucleosomal structure

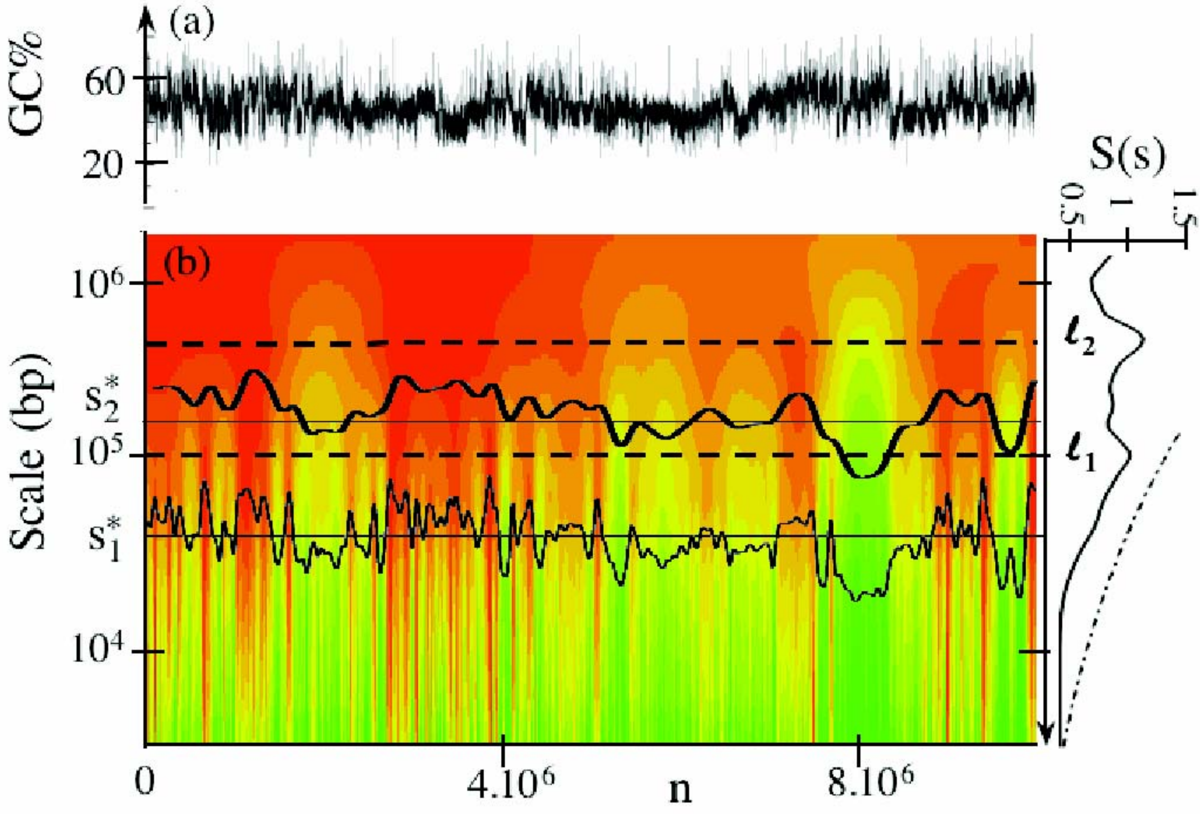
HIERARCHICAL STRUCTURE OF EUCARYOTIC DNA



LARGE SCALE REPRESENTATION OF GENOMIC SEQUENCES

Space-Scale Representation of the GC Content with a Smoothing Gaussian Filter

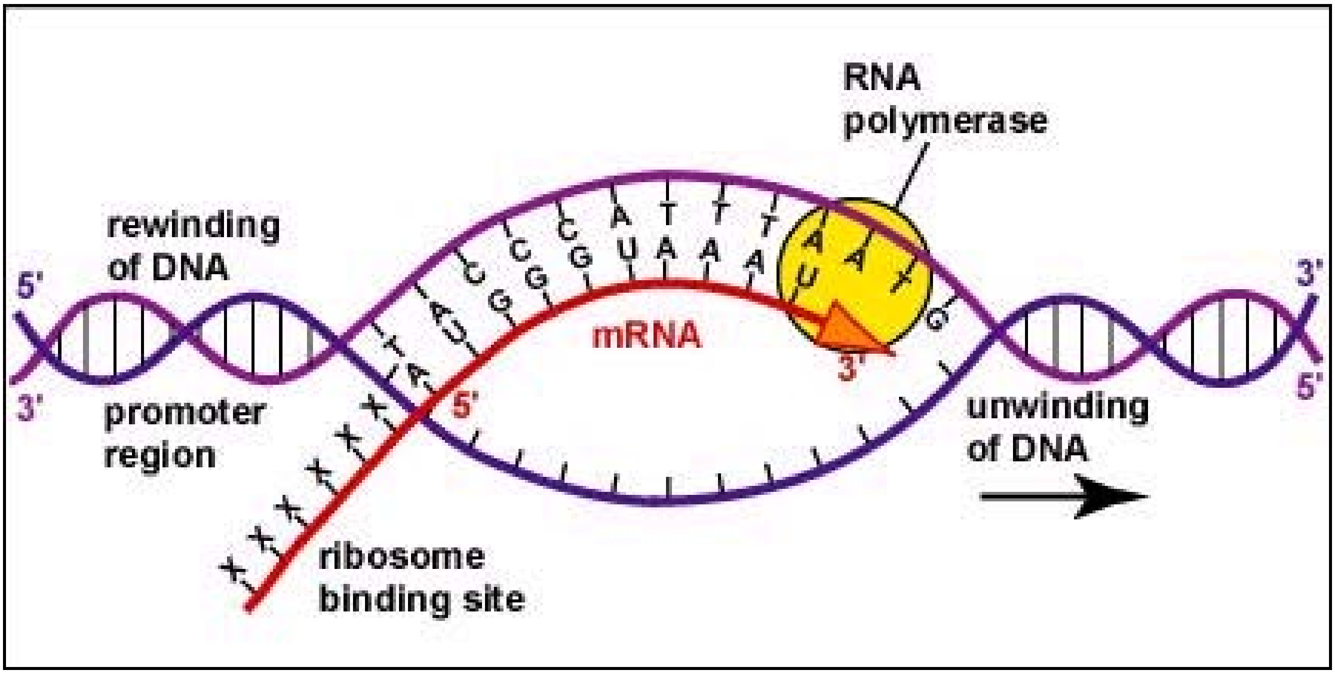
Chromosome 22 (Human)



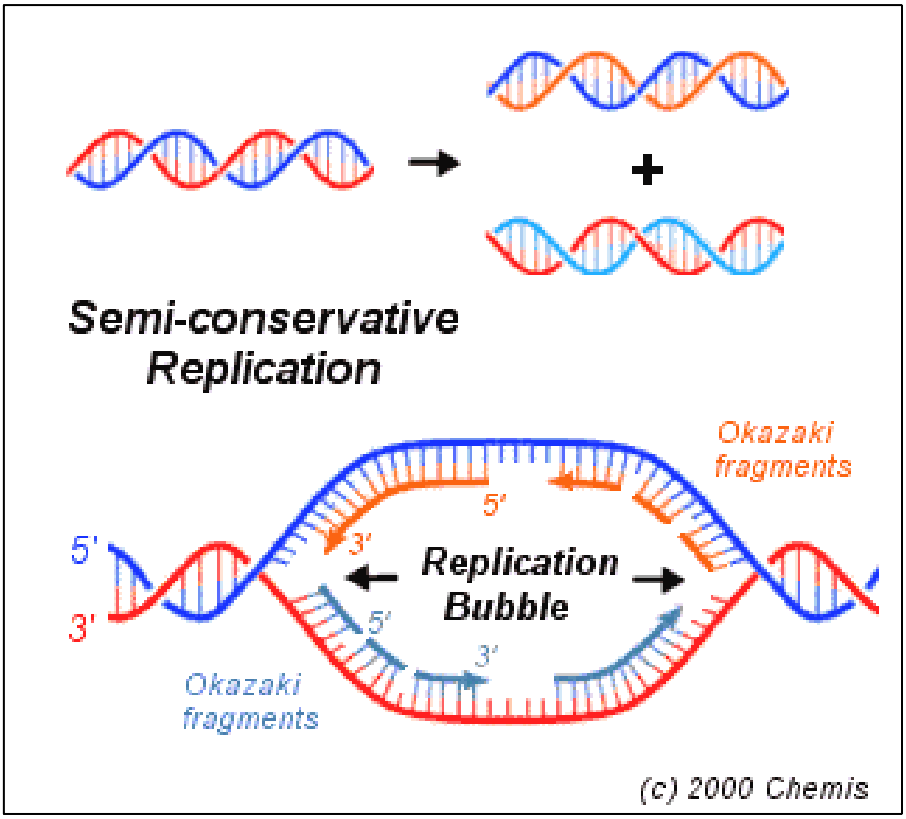
Filtering scales: $a_1^* = 40\text{kb}$, $a_2^* = 160\text{kb}$

Space-scale content: $S(a) = \sum_n |T_{\psi_M}(n,a)|$,
where ψ_M is the Morlet wavelet

Transcription



Replication



Opening of the double helix with a different environment for each strand => asymmetrical process

Symmetrical properties of the strands:
‘‘Parity Rule type 2’’

$$[A] = [T] \quad \& \quad [G] = [C]$$

in each strand

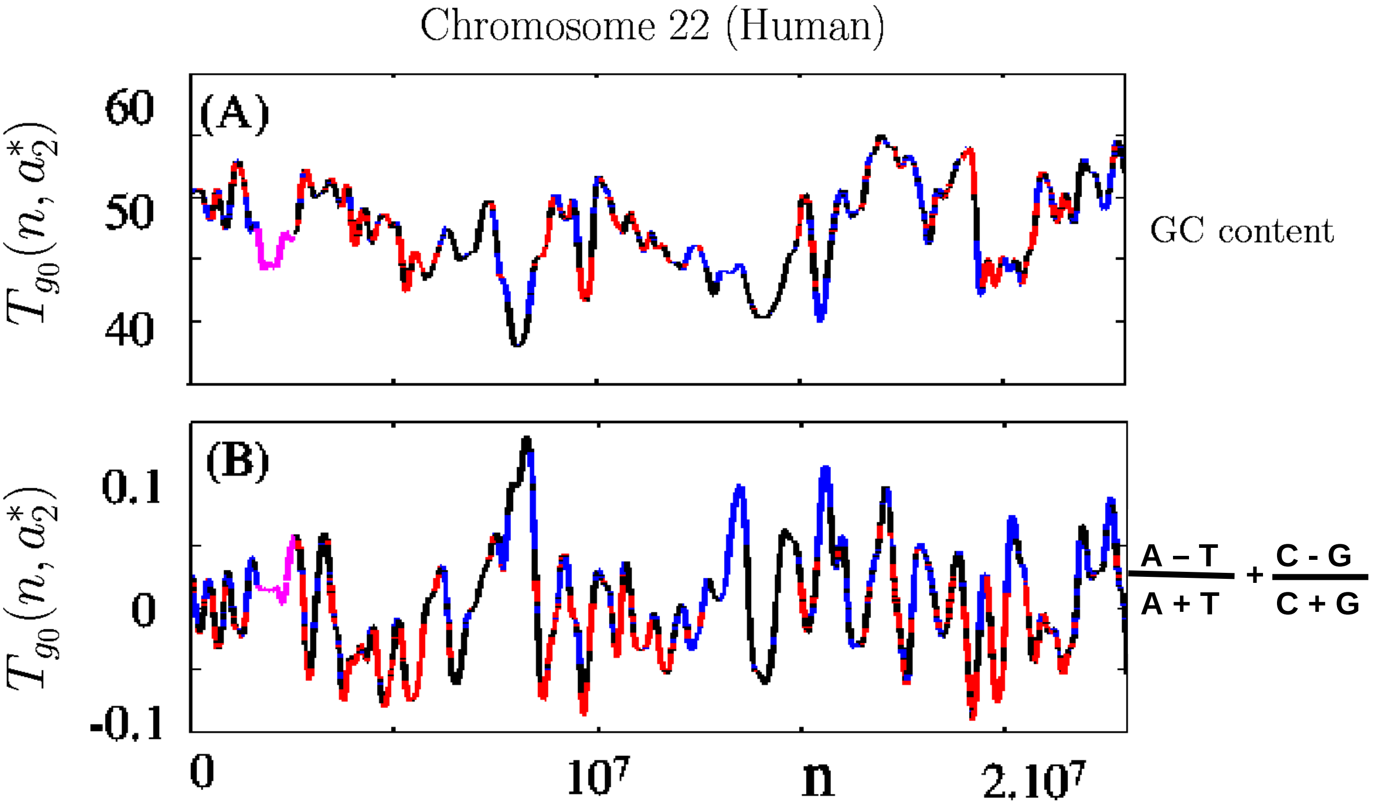
Deviations from this property estimated by the compositional skews

$$S_{CG} = \frac{[C] - [G]}{[C] + [G]}$$

$$S_{AT} = \frac{[A] - [T]}{[A] + [T]}$$

Compositional skew due to local biases in a strand in the course of biological mechanisms

Strand Compositional Asymmetry



–sense genes

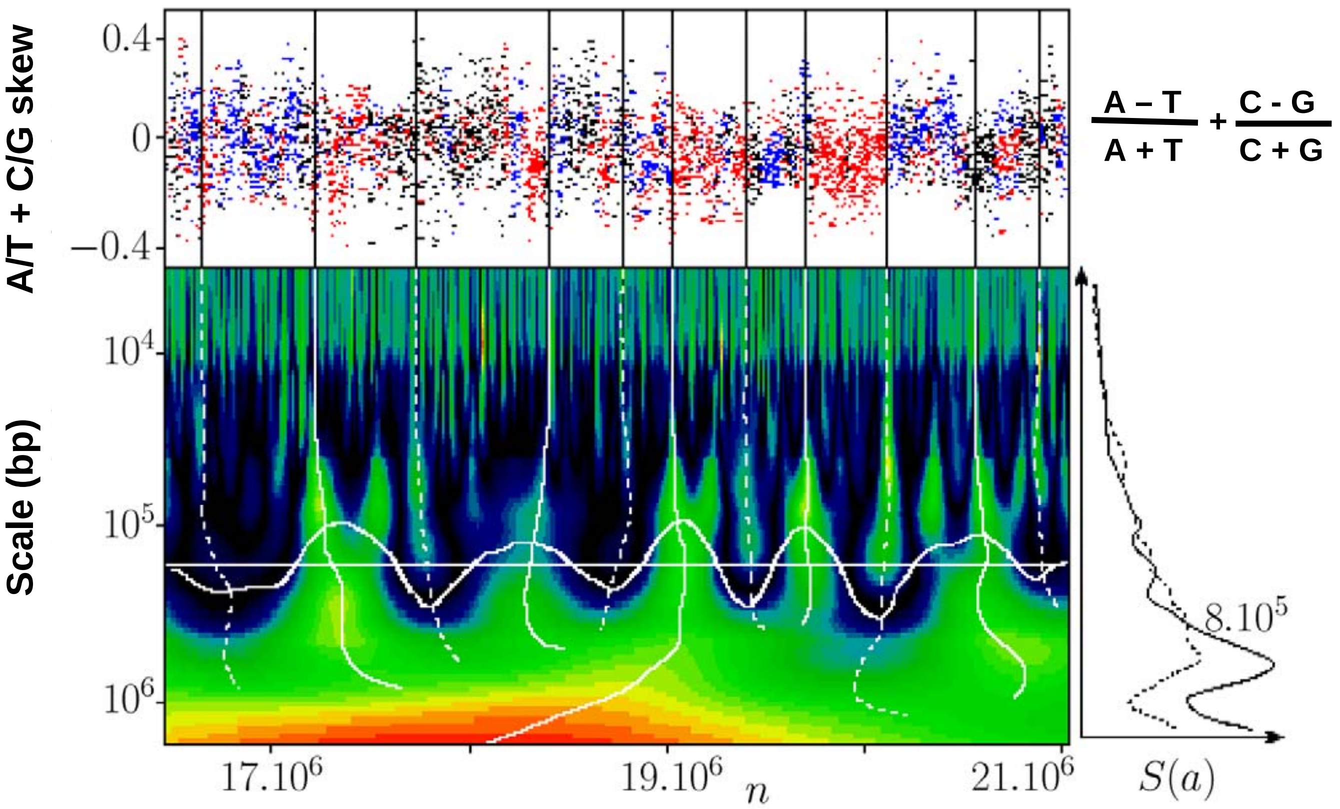
–anti-sense genes

–non-coding sequences

Filtering scales: $a_1^* = 40\text{kb}$, $a_2^* = 160\text{kb}$

A wavelet based methodology to detect gene clusters

Chromosome 22 (Human)



$$\frac{A-T}{A+T}+\frac{C-G}{C+G}$$

Analyzing wavelet: $g^{(n)}(x)=\frac{1}{\sqrt{2\pi}}\frac{d^n e^{-x^2/2}}{dx^n}$

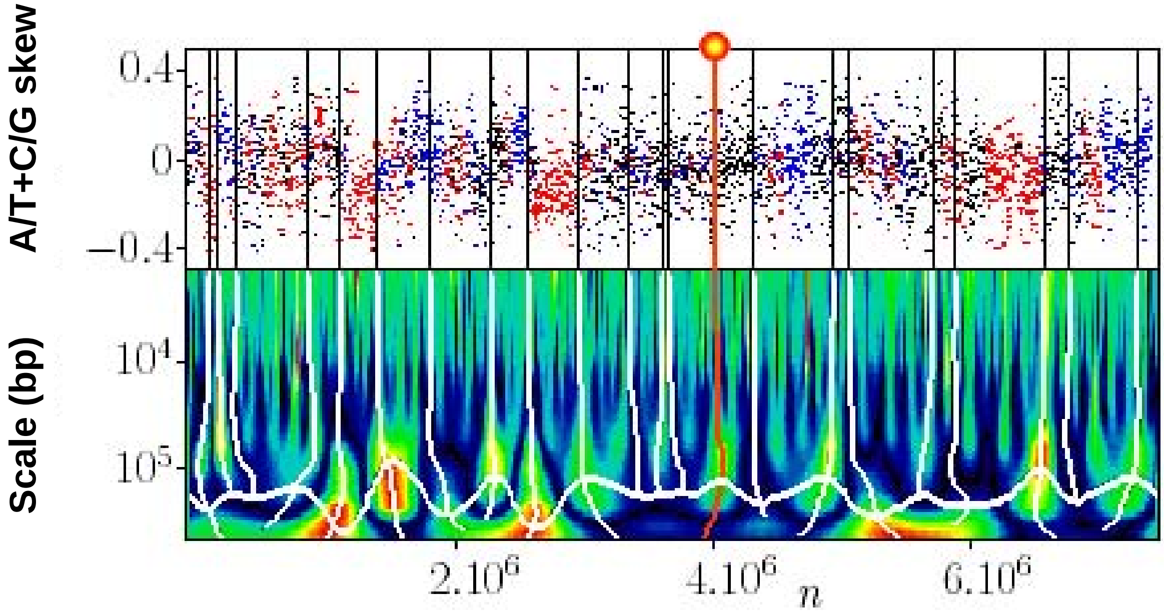
$$T_{g^{(n)}}(b,a)=\frac{1}{a}\int f(x)\; g^{(n)}\left(\frac{x-b}{a}\right)\,dx=\frac{d^n}{db^n}T_{g^{(0)}}(b,a)$$

A wavelet based methodology to detect replication origins

Experimentaly observed replication origin in the human genome

Globin: 4008 kb **Chromosome 11**

Predicted RO : 4009 kb



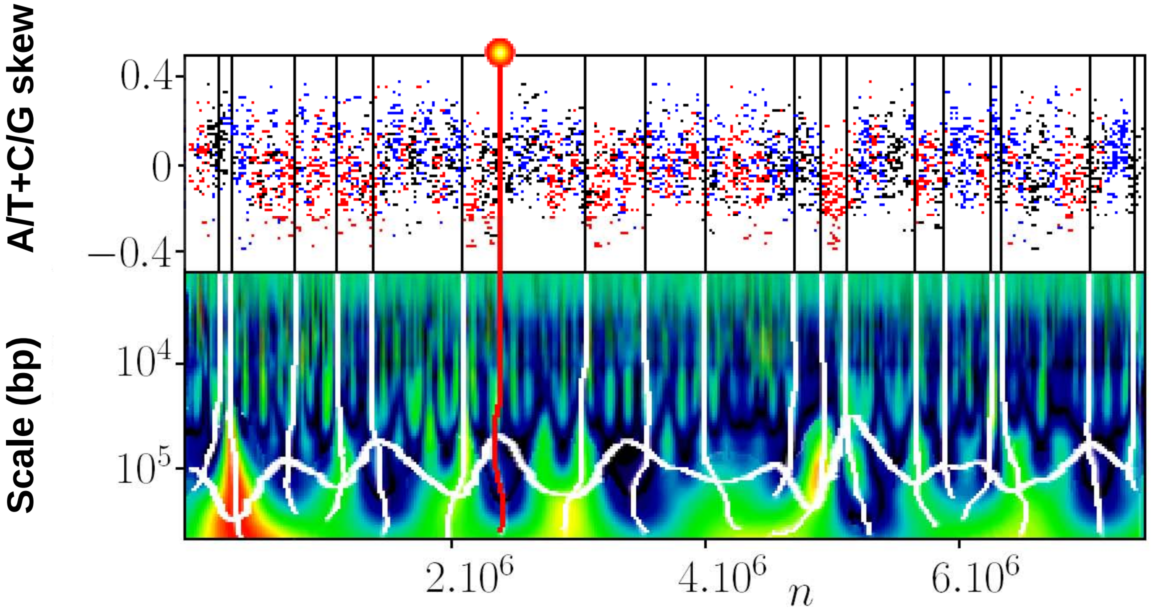
$$\text{Skew :} \quad \frac{A - T}{A + T} + \frac{C - G}{C + G}$$

A wavelet based methodology to detect replication origins

Experimentaly observed replication origin in the human genome

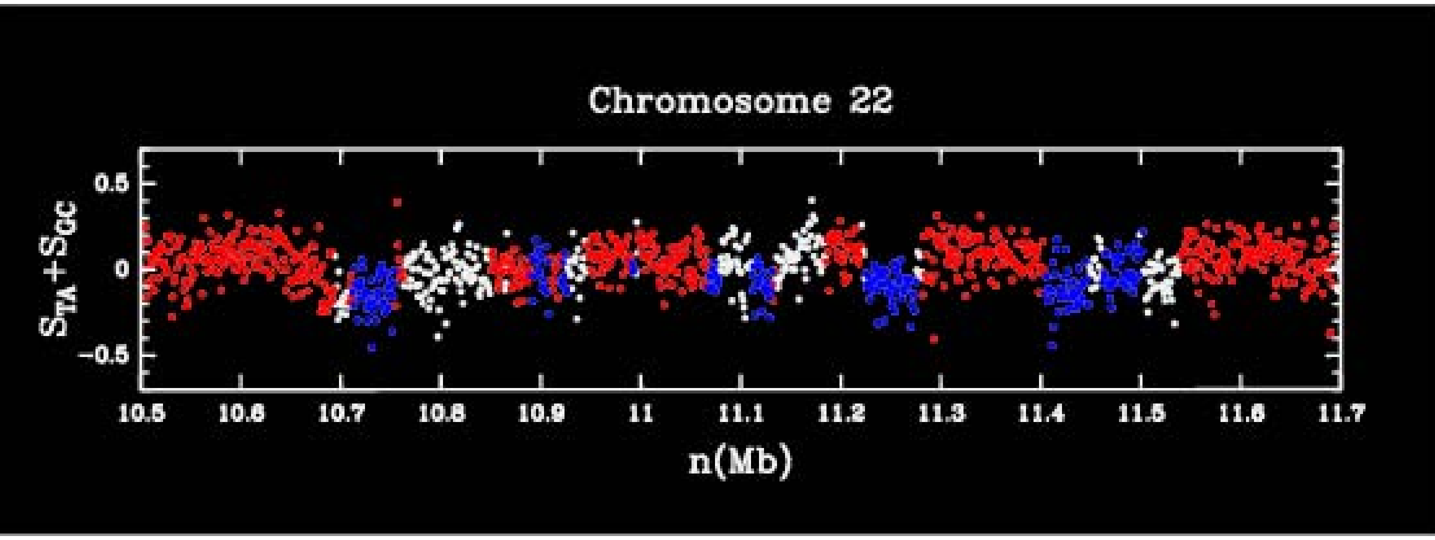
Lamin B2: 2368 kb **Chromosome 19**

Predicted RO : 2365 kb

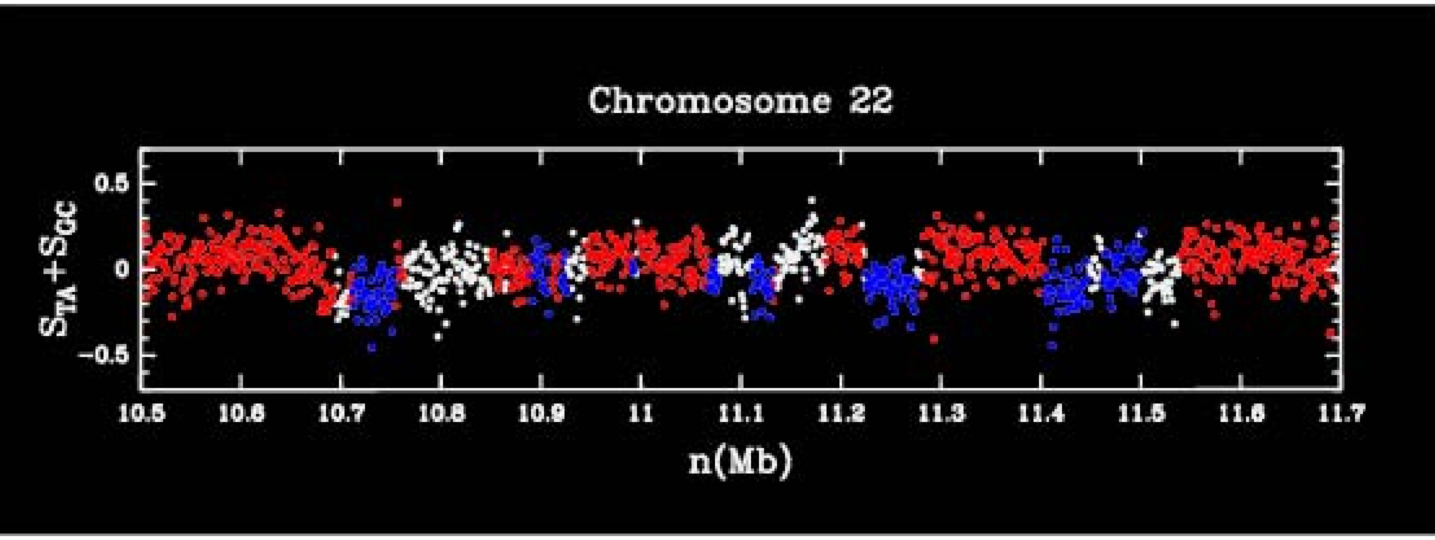


$$\text{Skew :} \quad \frac{A - T}{A + T} + \frac{C - G}{C + G}$$

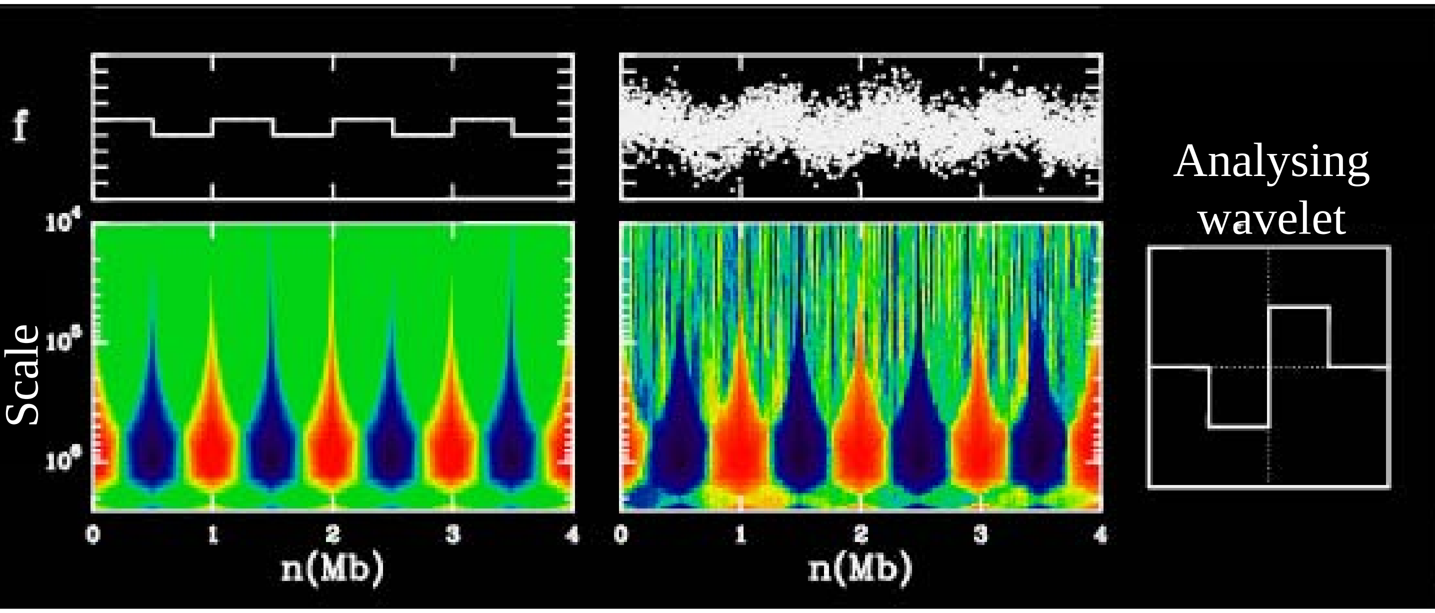
Transcription bias



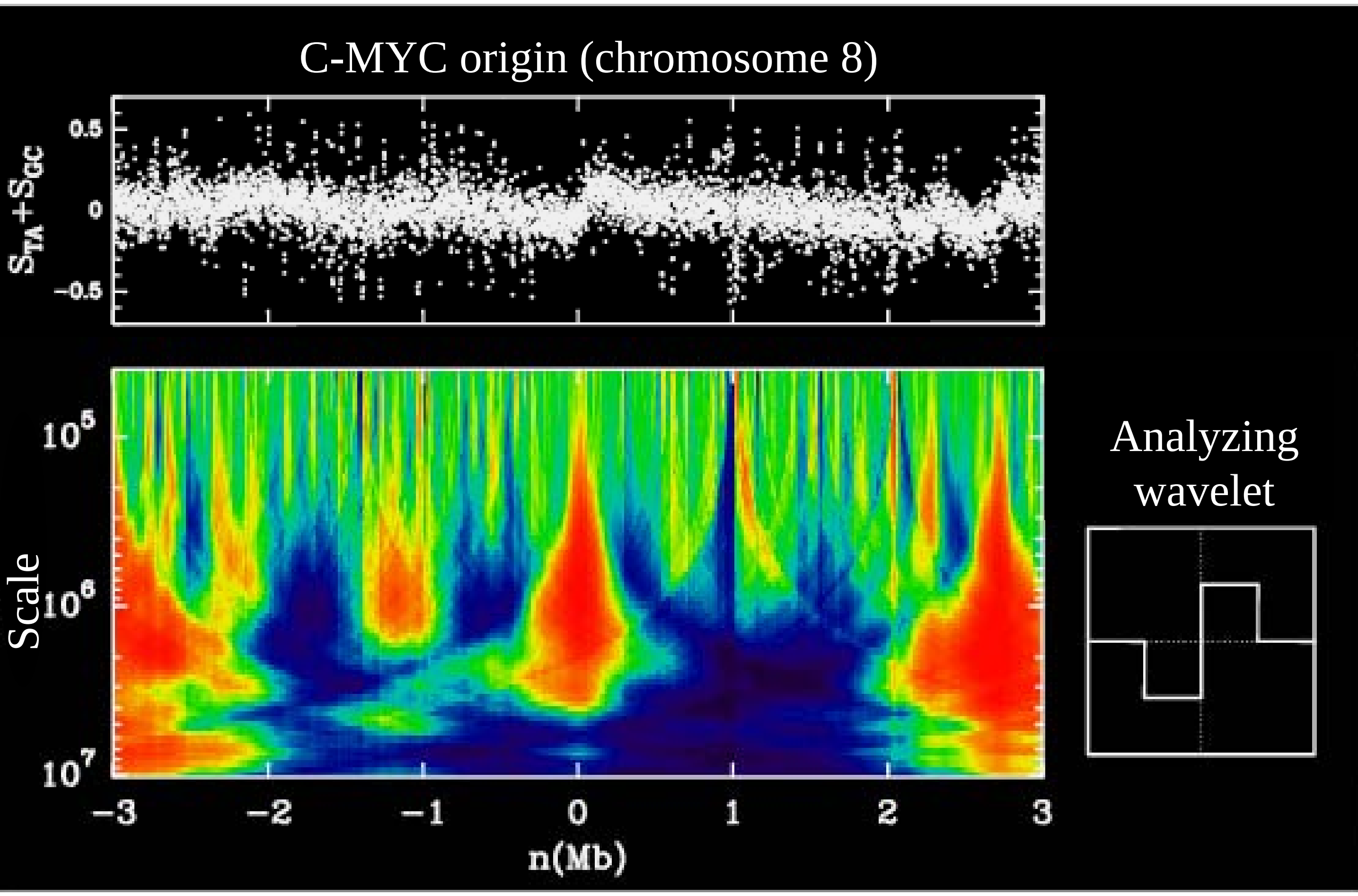
Transcription bias



Detecting discontinuities using the wavelet transform



Application to a known human replication origin

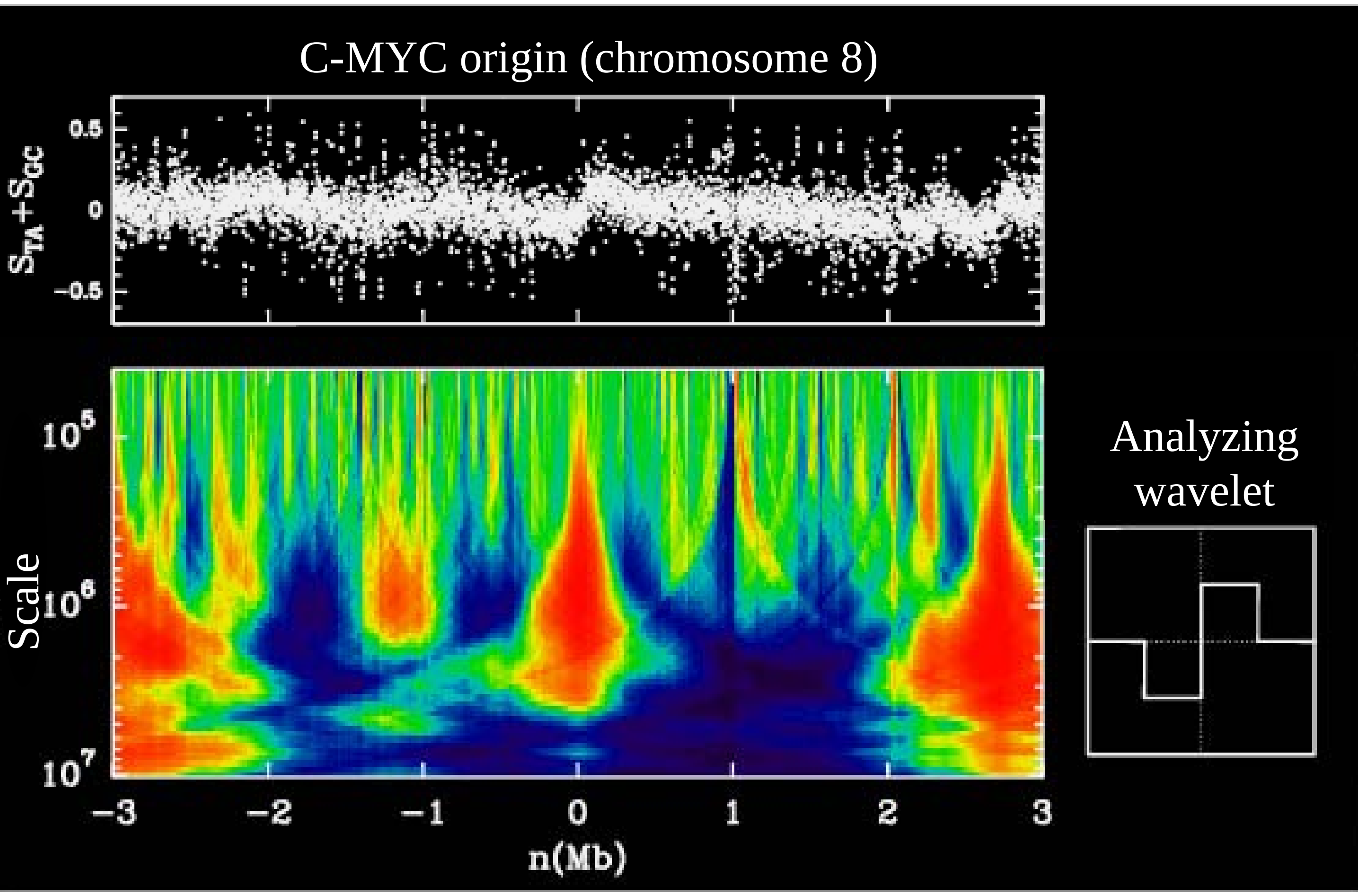


C-MYC origin (chromosome 8)

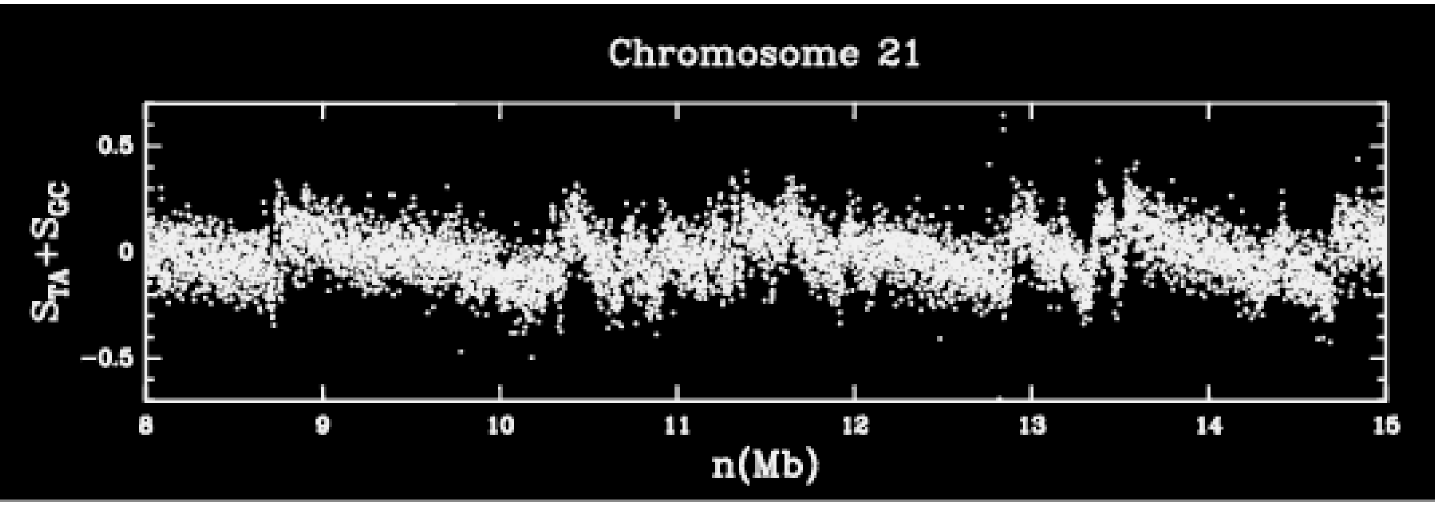
Analyzing wavelet

First evidence of a replication bias in human DNA

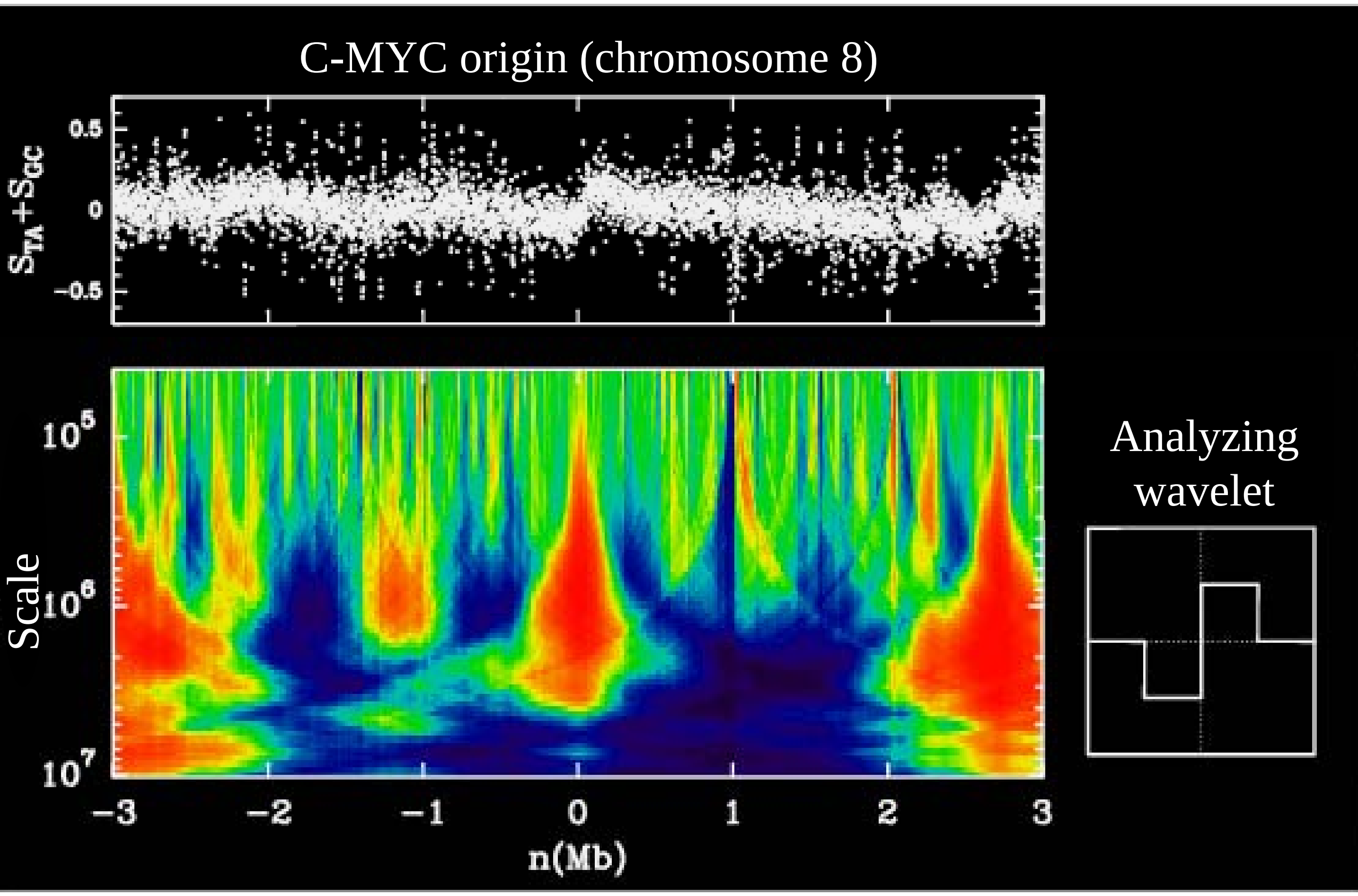
Application to a known human replication origin



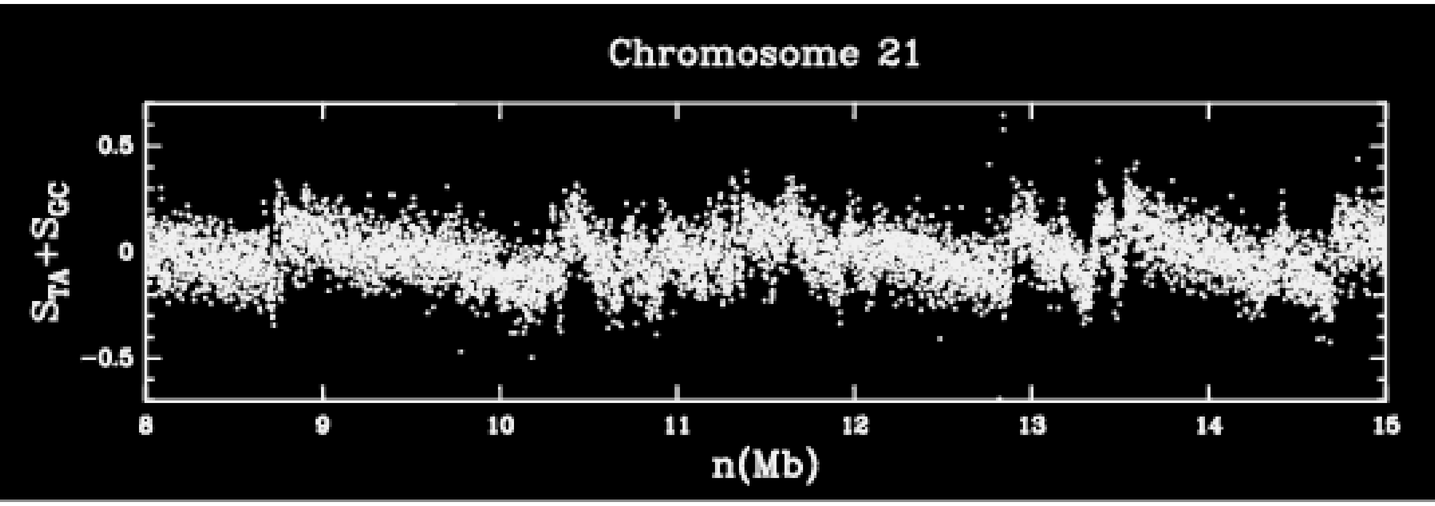
First evidence of a replication bias in human DNA



Application to a known human replication origin

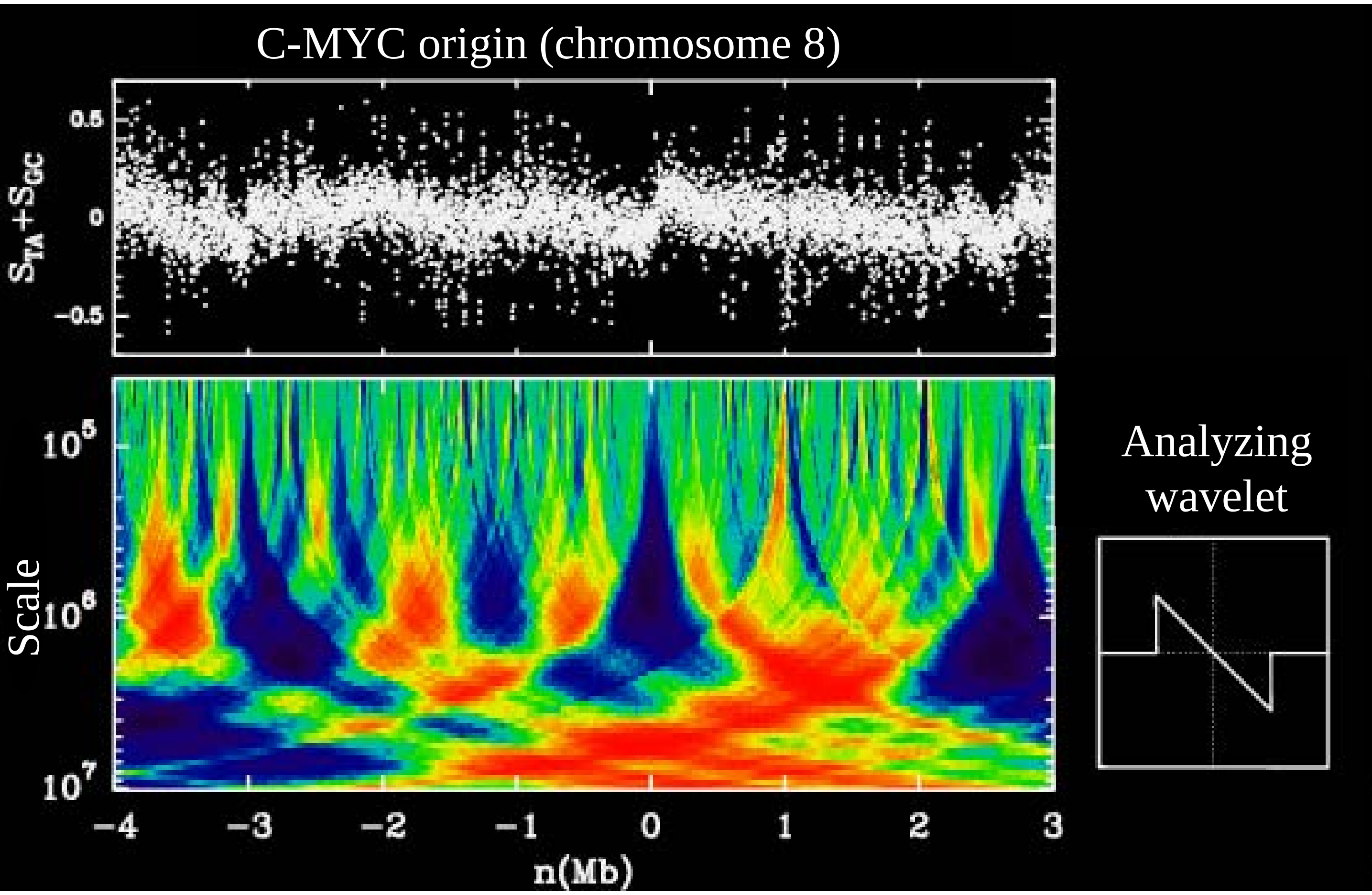


First evidence of a replication bias in human DNA



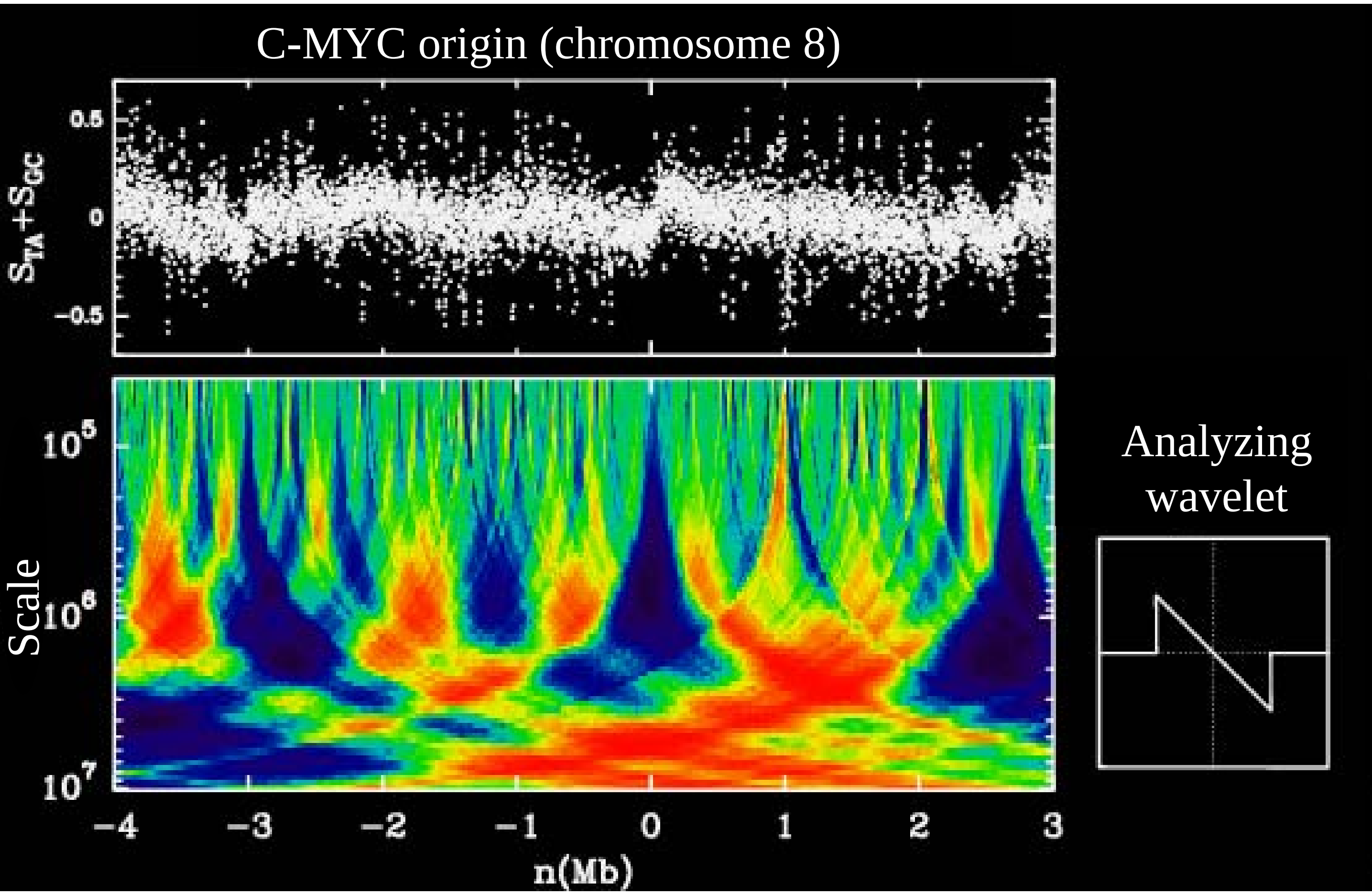
Our model : **well defined replication origins, separated by diffuse terminuses**

Profile detection using an analyzing wavelet adapted to the shape of replicons

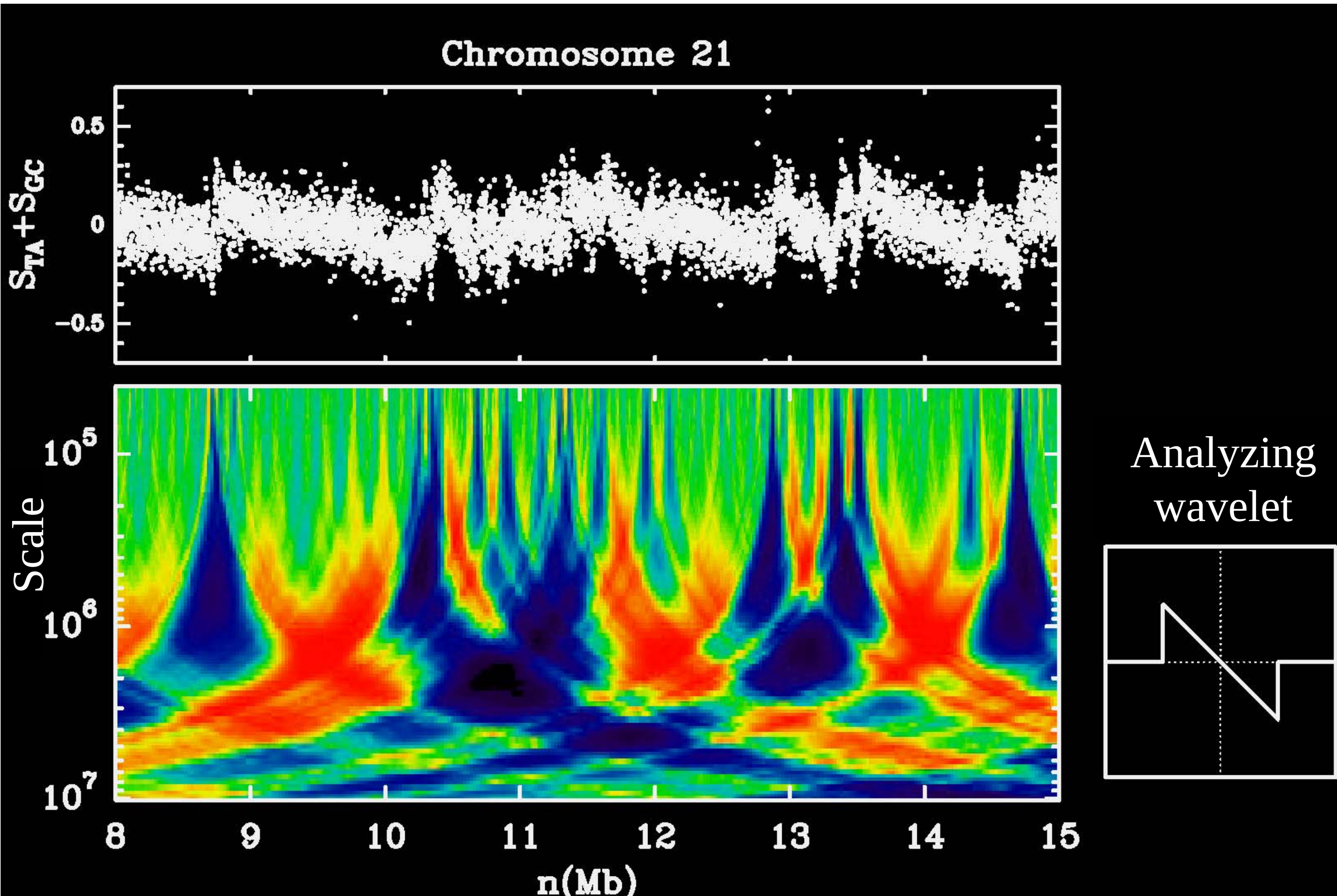


C-MYC origin (chromosome 8)

Profile detection using an analyzing wavelet adapted to the shape of replicons

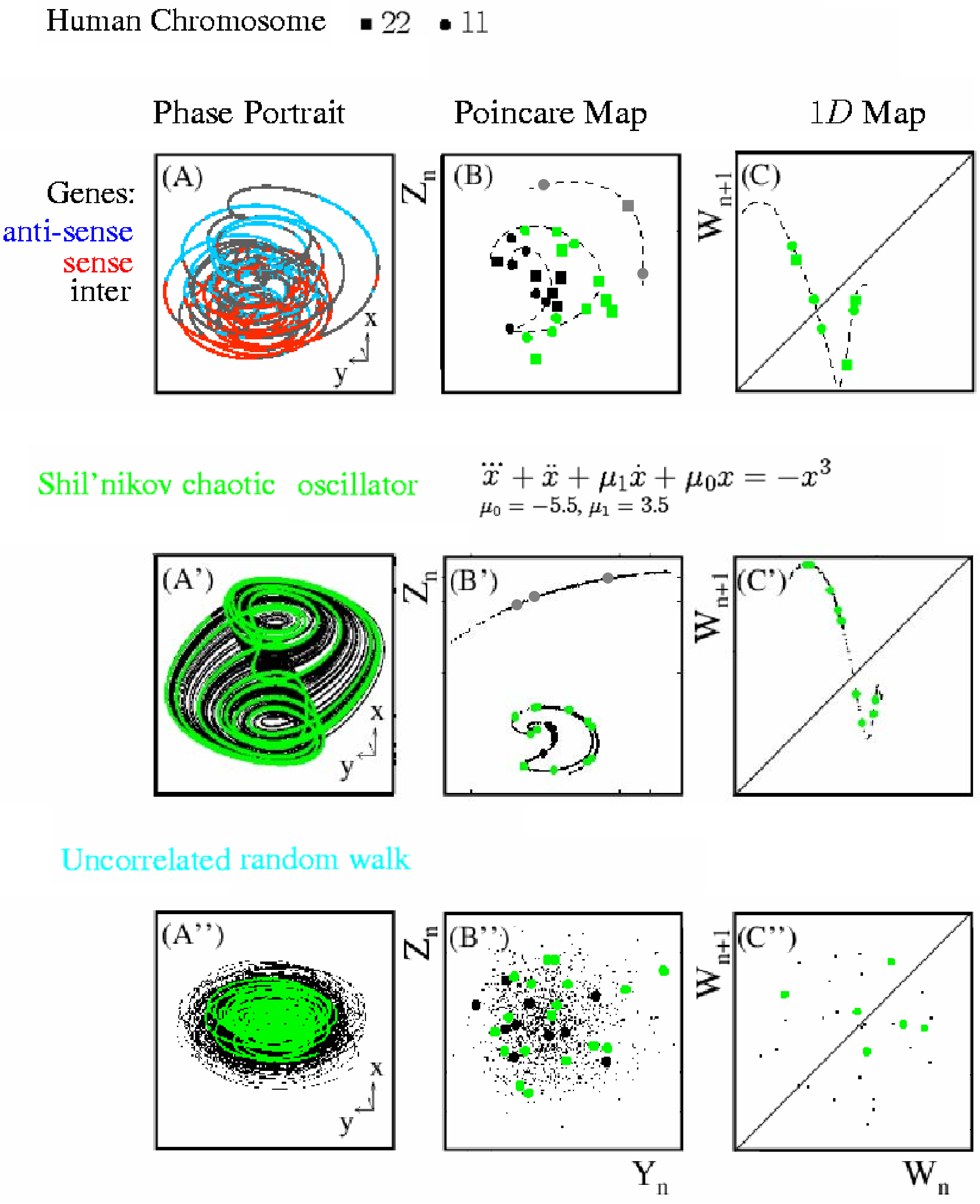


C-MYC origin (chromosome 8)



Chromosome 21

Deterministic Chaos in DNA Sequences


Human Chromosome ■ 22 ● 11

Phase Portrait

Poincare Map

$1D$ Map

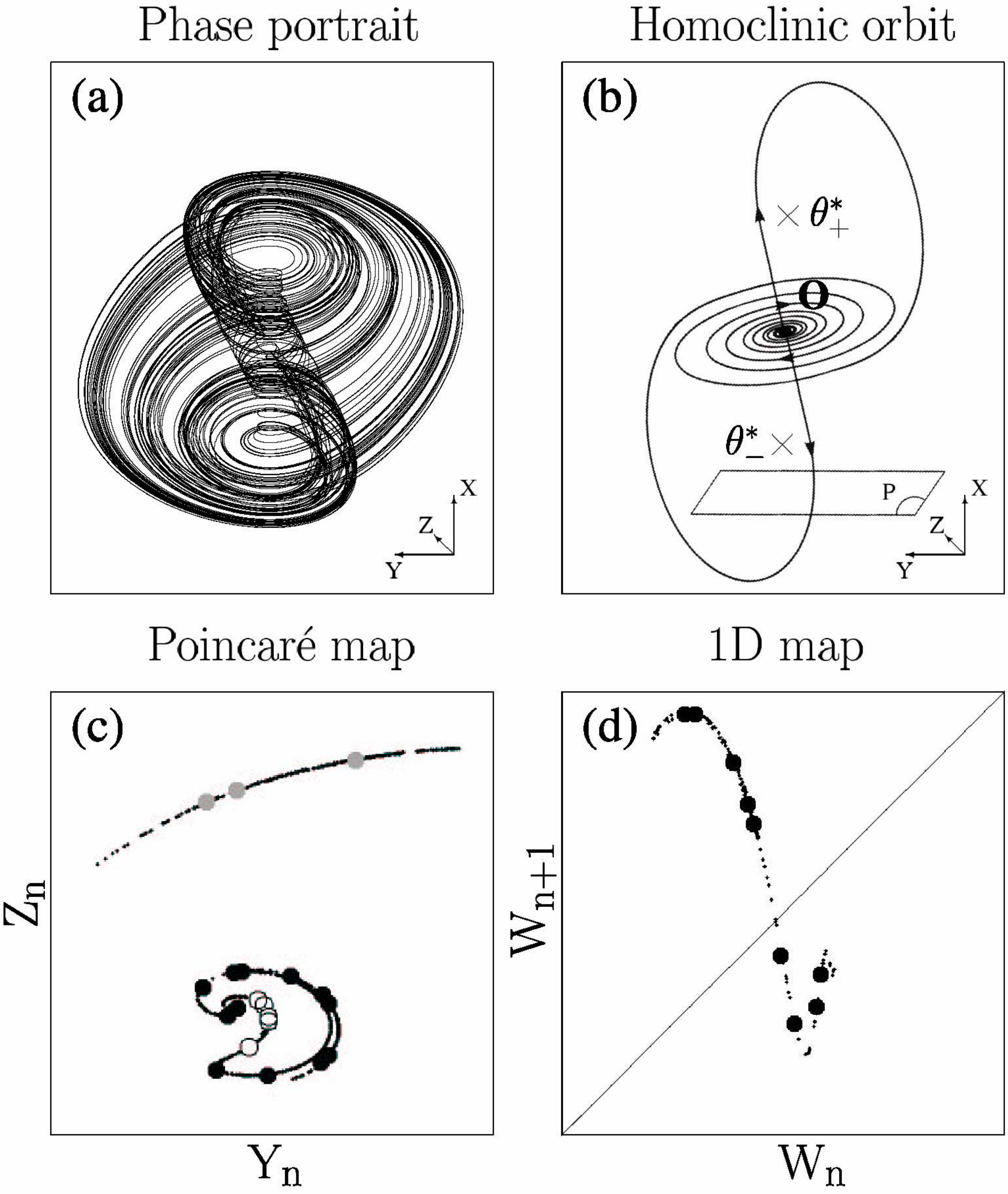
Genes:
anti-sense
sense
inter

Shil'nikov chaotic oscillator

$$\dddot{x} + \ddot{x} + \mu_1 \dot{x} + \mu_0 x = -x^3$$
$$\mu_0 = -5.5,\ \mu_1 = 3.5$$

Uncorrelated random walk

SHIL’NIKOV HOMOCLINIC CHAOS



Phase portrait

Homoclinic orbit

Poincaré map

1D map

LYAPUNOV EXPONENTS

$S_{AT} - S_{GC}$ skew profiles smoothed at scale 160 kb

	d				
	3	4	5	6	7
Chromosome 11 (24Mb) (NT_033899.3)	12.6	8.9	6.1	6.9	7.9
Chromosome 14 (68Mb) (NT_026437.9)	15.0	10.2	8.8	8.7	10.4
Chromosome 21 (29Mb) (NT_011512.7)	12.2	8.7	7.4	8.6	11.3
Chromosome 22 (23Mb) (NT_011520.8)	12.5	8.1	6.3	5.8	7.2
Shil'nikov strange attractor (30Mb)	4.2	5.6	6.5	7.3	7.1

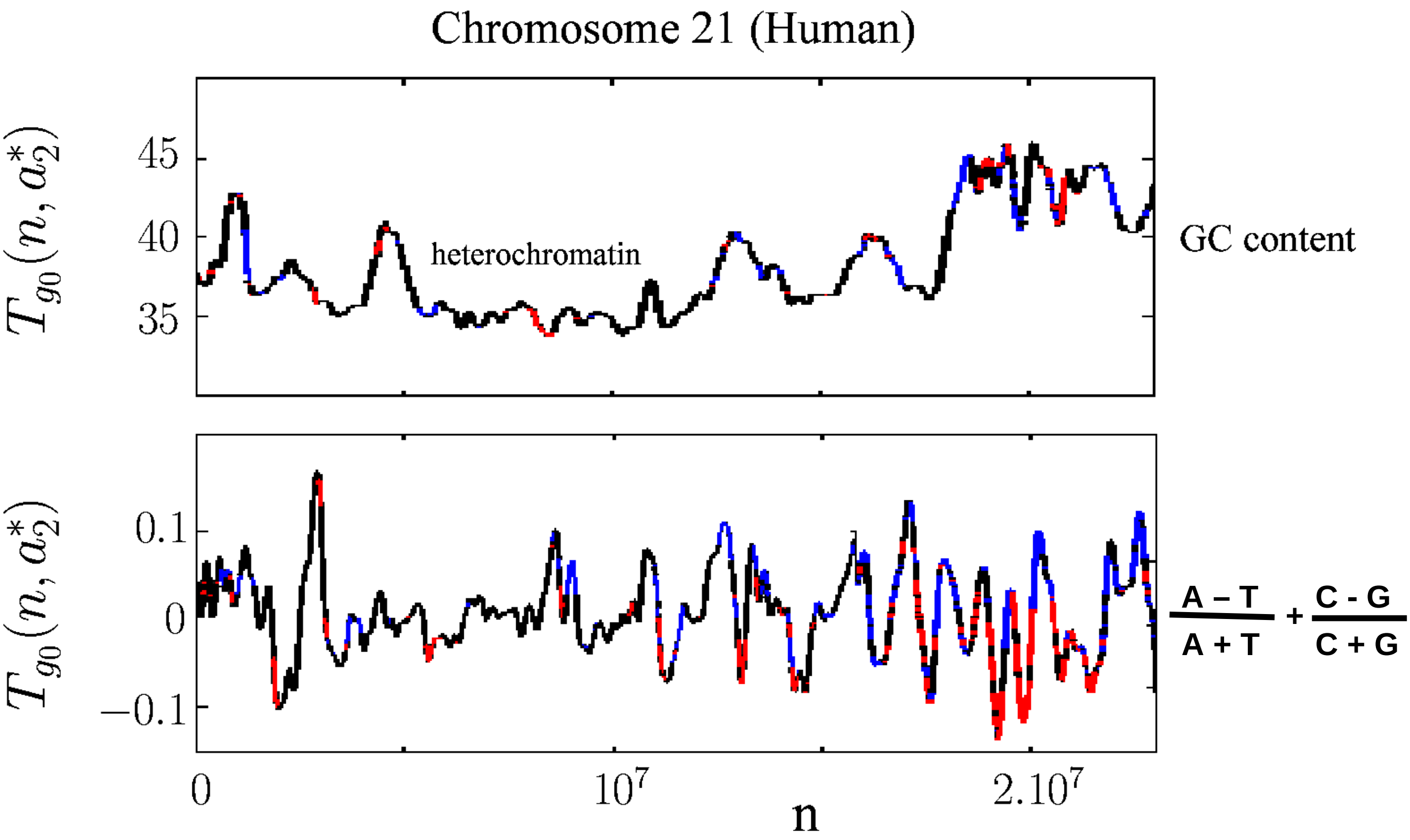
Computation of the largest Lyapunov exponent ($\times 10^3$) using the TISEAN package for a time delay $\tau = 60$ kb and an embedding dimension d.

Equation of non-linear oscillator which displays homoclinic chaos of Shil'nikov's type:

$$\dddot{\theta} + \mu_2 \ddot{\theta} + \mu_1 \dot{\theta} + \mu_0 \theta + k\theta^3 = 0$$

θ and t were rescaled so that the chaotic trajectory displays similar amplitude and characteristic frequencies as the skew oscillatory profiles.

Strand Compositional Asymmetry



$$\frac{A - T}{A + T} + \frac{C - G}{C + G}$$

–sense genes

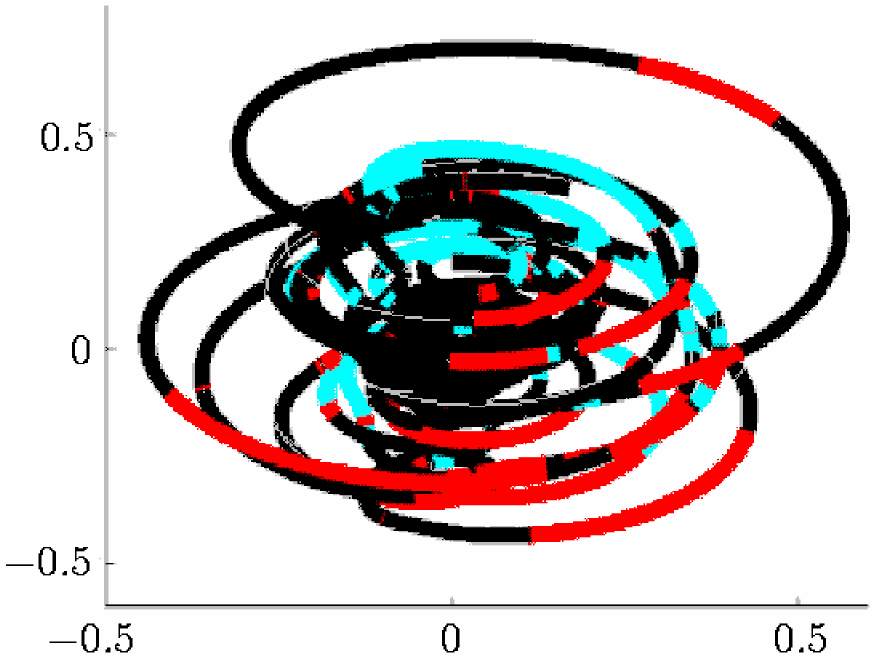
–anti-sense genes

–non-coding sequences

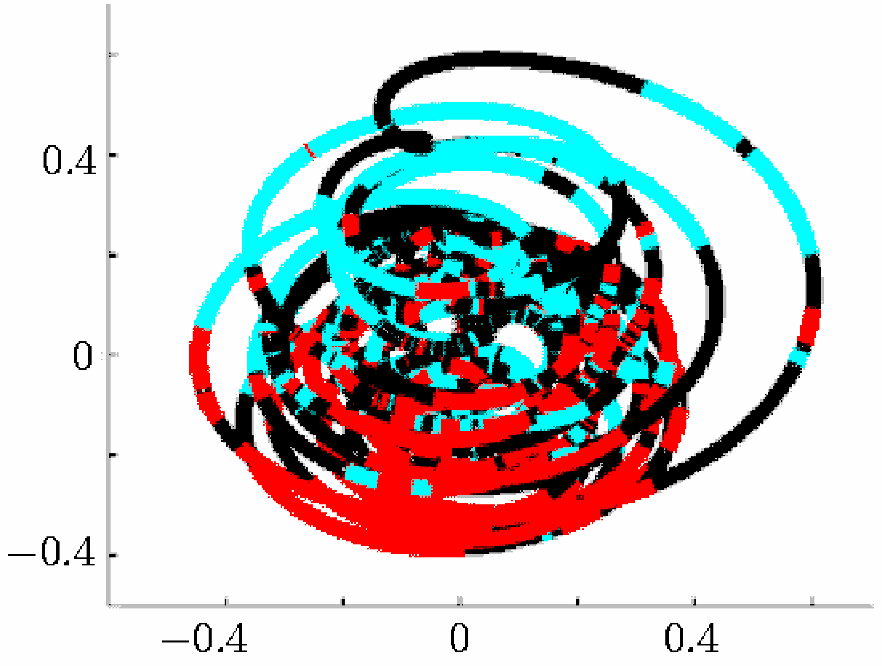
Filtering scales: $a_1^* = 40\text{kb}$, $a_2^* = 160\text{kb}$

Phase Portrait Representation of AT+CG skew

Chromosome 21



Chromosome 22



Filtering scale: $a_2^* = 160\text{kb}$

REFERENCES

Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes.

M. TOUCHON, A. ARNEODO, Y. D'AUBENTON-CARAFA & C. THERMES, Nucleic Acids Res. (2004), to appear

Low Frequency rhythms in human DNA sequences : a key to the organization of gene location and orientation?

S. NICOLAY, F. ARGOUL, M. TOUCHON, Y. D'AUBENTON-CARAFA, C. THERMES & A. ARNEODO, Phys. Rev. Lett. (2004), to appear

From scale invariance to deterministic chaos in DNA sequences : towards a deterministic description of gene organization in the human genome

S. NICOLAY, E.B. BRODIE OF BRODIE, M. TOUCHON, Y. D'AUBENTON-CARAFA, C. THERMES & A. ARNEODO, Physica A (2004), to appear

Transcription-coupled TA and GC strand asymmetries in the human genome.

M. TOUCHON, S. NICOLAY, A. ARNEODO, Y. D'AUBENTON-CARAFA & C. THERMES, FEBS Letters **555**, 579 (2003)

Long-range correlations between DNA bending sites : relation to the structure and dynamics of nucleosomes.

B. AUDIT, C.VAILLANT, A. ARNEODO, Y. D'AUBENTON-CARAFA & C. THERMES, J. Mol. Biol. **316**, 903 (2002)

Long-range correlations in genomic DNA : a signature of the nucleosomal structure.

B. AUDIT, C. THERMES, C. VAILLANT, Y. D'AUBENTON-CARAFA, J.F. MUZY & A. ARNEODO, Phys. Rev. Lett. **86**, 2471 (2001)

Nucleotide composition effects on the long-range correlations in human genes.

A. ARNEODO, Y. D'AUBENTON-CARAFA, B. AUDIT, E. BACRY, J.F. MUZY & C. THERMES, Eur. Phys. J. **B1**, 259 (1998)

Wavelet based fractal analysis of DNA sequences.

A. ARNEODO, Y. D'AUBENTON-CARAFA, E. BACRY, P.V. GRAVES, J.F. MUZY & C. THERMES, Physica **96 D**, 291 (1996)

Characterizing long-range correlations in DNA sequences from wavelet analysis.

A. ARNEODO, E. BACRY, P.V. GRAVES & J.F. MUZY, Phys. Rev. Lett. **74**, 3293 (1995)